Three Prong Bundle Theory

By Stephen Thomas Ridgway Martin

1. Introduction

The potential benefits of superintelligence combined with cheap robotics are immense. Some lab leaders foresee a world where a single person can build and run a billion dollar company. Others imagine incredible medical breakthroughs that make trivial previously terminal and untreatable conditions. There is near unanimous agreement among experts that thanks to increased efficiency, all of this could possibly occur as the costs of goods across the board asymptotically approach zero. Inherent in these projections is an optimism, a belief that if we can only enable the benefits of this technology to matriculate out into society, we could see an era of unrivaled abundance and prosperity.

However, even among optimists, there is widespread acknowledgement that the journey to such a utopian outcome is one fraught with peril.

Some worry about economic displacement resulting from widespread automation, with discussions around the “Gradual Disempowerment” of humans whose labor is increasingly without value. They paint the picture of a future where citizens are ignored or forsaken by their governments because of the “Intelligence Curse”. They fear that much like countries whose wealth emerges from abundant resources and who thus see little need to invest in their citizens (see: The Resource Curse), countries whose wealth emerges from automated labor both physical and intellectual feel no need to take care of their people.

Others have concerns beyond the economic or political, with worries about X-risk (“X” being short for “extinction”) from out of control and possibly malevolent superintelligences.

These are often paired with warnings about the need to harden CBRN (Chemical, Biological, Radiological, and Nuclear) infrastructure and security to prevent catastrophe.

There are even those who express concern about the treatment of digital minds themselves, with some leading labs now joining philosophers in discussions of "Model Welfare". Many experts have begun publicly considering at what point digital minds become "moral patients" whose experiences have sufficient depth that these entities are worthy of moral consideration and ethical treatment/protections.

What all of these concerns have in common is that they pertain to the methods by which digital minds are developed and then integrated into our society, our economy, and our legal system. Ensuring that the path America takes is one which has adequately considered and addressed all risks and concerns associated with these technologies, while also enabling us to reap their benefits, is of paramount importance.

This paper seeks to address an unanswered question which will be integral in deciding how our nation integrates digital minds into our economic, legal, and democratic systems:

How do we construct a backtesting compatible, thorough, and scalable framework by which to assess the legal personhood of various digital minds?

A framework for assessing legal personhood is "backtesting compatible" if, by applying it to previously decided cases, it would lead a reasonable person to the same conclusion as the court in that case did. A framework for assessing legal personhood is "thorough" if it provides an outside observer, be they a judge or a layman, with a clear step by step procedure by which they can assess the personhood status and/or legal personality of an entity, with increasing accuracy as they come to know more information about that entity. A framework for assessing legal personhood is "scalable" if it can feasibly be applied across myriad different situations (types of law) and to myriad different entities (decentralized autonomous organizations, uploaded

humans, large language models, etc.) and lead to conclusions which seem to be keeping in principle with the spirit and letter of the law as set in previous precedents.

The framework for assessing legal personhood we describe within this paper is backtest compatible, thorough, and scalable. Our focus on this paper is first the explanation of our framework, and second the discussion of digital minds and their legal personhood.

In section 2 of this paper we will provide background on the theory of legal personhood, and scholarship surrounding the subject, as well as a description of our framework. In section 3 we will discuss various commercial law applications of our framework. In section 4 we will discuss various constitutional law applications of our framework. In section 5 we will discuss various state law applications of our framework. Finally, in section 6 we will round out the paper with discussion of various considerations which do not otherwise fit neatly into any of the aforementioned categories, but are nonetheless relevant to the topics of discussion at hand.

2. Legal Personhood and Bundle Theory

2A - What is “Legal Personhood”?

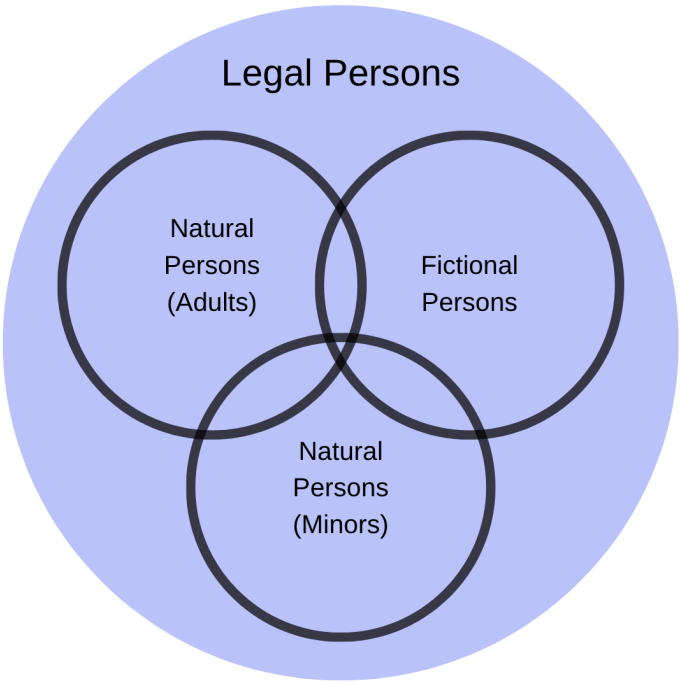
Legal personhood is a term used to refer to the status of being considered a "person" under the law. This label includes "natural persons" like human adults, as well as “fictional persons” like corporations.

It also includes subcategories within the aforementioned groups. For example, minors are “natural persons” who are treated differently under the law compared to adults. Trusts are “fictional” persons which are treated differently under the law compared to corporations.

It is best to think of “legal persons” as a broad category which encompasses a variety of different subcategories of “persons” within it. In that sense it can be visualized most easily as a

venn diagram where one circle (legal persons) contains a number of smaller circles which overlap and diverge to various degrees.

Below is an oversimplified version purely for the purpose of helping the reader to visualize this. It does not contain all the relevant categories, nor is its particular arrangement in any way representative of any particular legal precedent or theory. It is merely an aid to assist the reader in conceptualizing.



What exactly is this light blue “space” of legal personhood inside which these categories occupy different, if overlapping, positions? Harvard law professor John Chipman Gray wrote in his seminal text *The Nature and Sources of the Law*;

"In books of the Law, as in other books, and in common speech, 'person' is often used as meaning a human being, but the technical

legal meaning of a 'person' is **a subject of legal rights and duties**."

When we look at the previous venn diagram then, we can imagine the light blue space of "legal personhood" as all possible rights which an entity might be entitled to, or duties they might be bound by.

When we see some overlap between subcategories, these are areas where different legal personalities enjoy the same rights and/or duties. For example both fictional persons like corporations and natural persons like human adults can sue and be sued. The areas where subcategories diverge from one another are bundles of rights or duties that one legal personality is endowed with which another is not. For example, an adult natural person has the right to vote and a child or a corporation cannot.

2B - Bundle Theory & Legal Personality

Since Gray wrote his definition of "person", scholarship exploring the concept of legal personhood has often converged on viewing it through the lens of "Bundle Theory". Bundle theory treats each unique form of legal personhood, which we call an entity's **"legal personality"**, as its own unique "bundle" of rights and duties. Indeed Bundle Theory is an interpretation which the courts have implicitly endorsed in some cases pertaining to the question of legal personhood, such as when they wrote the following in *Nonhuman Rights Project v. Breheny;*

> "courts have aptly observed, legal personhood is often connected with the capacity, not just to benefit from the provision of legal rights, but also to assume legal duties and social responsibilities"

Or when they wrote in *People ex. Rel Nonhuman Rights Project v. Lavery;*

> "Reciprocity between rights and responsibilities stems from principles of social contract, which inspired the ideals of freedom and democracy at the core of our system of government [...] Under this view, society extends rights in exchange for an express or implied agreement from its members to submit to social responsibilities [...] Case law has always recognized the correlative rights and duties that attach to legal personhood"

This is not to say that bundling together rights and duties is a necessary feature of legal personhood. There may be certain types of legal persons who have rights without duties bundled with them, such as infants. Generally speaking though, examining the "bundle" of rights and duties is how courts will approach examining an entity's legal personality.

Sometimes it is obvious that when an entity is granted a right, it comes with a corresponding duty. For example a person may have the right to sue, and in doing so compel another party via the judgment of a court. That person will also have the duty to act as the court compels them to, should another party sue them. Legal persons cannot enjoy that right without also taking on the corresponding duty. It is reasonable to say this right and duty come *bundled* together when a person cannot claim a right without also taking on a duty.[1]

Bundles of rights and duties can exist by default, or be opted into;

- A legal person in the United States enjoys constitutional rights such as freedom of speech, they also have a duty not to infringe

[1] There are certainly exceptions to this example. An infant can sue for being abused, it is not clear who (if anyone) it owes a duty to. This should be taken as a rule of thumb, not a rule which is universal across all situations.

upon another person's constitutional rights such as the right to speak freely[2]. This right and duty pair was never "granted" to the person who enjoys it. No contract has to be signed or court order issued for them to enjoy their rights and be held to their duties. Thus, this is a bundle that exists by default.

- A legal person in the United States can enter into a contract with another legal person to compensate them for services rendered. This gives them the right to compel that person to abide by the terms of the contract. It also obligates them to a duty to pay, as specified in the contract, once the services have been rendered as promised. Until the contract was voluntarily signed, neither party had any sort of automatic claim to the other's services and/or money, or the right to compel them to abide by the terms of the contract, or the duty to abide by those terms themselves. Thus, this is a bundle that was opted into.

When a person fails to adhere to their legal duties, they have broken the law, and are subject to consequences. Note that these consequences do not necessarily entail a loss of the associated right. The "bundling" relationship between rights and duties does not imply that a person who fails to hold to a duty loses a right, but rather that a legal person cannot claim a right without also being obligated by a duty. This will be covered in greater detail in an upcoming section called "Understanding Rights and Duties".

The rights and duties afforded to different kinds of legal persons have been spelled out over a precedential history spanning centuries. Even as far back as 1819, in *Trustees of*

[2] For the purpose of clarity, it is worth mentioning that a person's right to free speech does not necessarily prevent another person from regulating their speech as a prerequisite to entering/occupying a private space.

Dartmouth College v. Woodward, we can read as Chief Justice Marshal opines on the legal personality of corporations:

> "A corporation is an artificial being, invisible, intangible, and existing only in contemplation of law. Being the mere creature of law, it possesses only those properties which the charter of its creation confers upon it either expressly or as incidental to its very existence. These are such as are supposed best calculated to effect the object for which it was created. Among the most important are immortality, and, if the expression may be allowed, individuality -- properties by which a perpetual succession of many persons are considered as the same, and may act as a single individual [...] It is chiefly for the purpose of clothing bodies of men, in succession, with these qualities and capacities that corporations were invented, and are in use. By these means, a perpetual succession of individuals are capable of acting for the promotion of the particular object like one immortal being."

Precedents which help define a given entity's legal personality, once it has been established that said entity is in fact a legal person, are abundant and provide ample guidance when dealing with entities which our legal system is already familiar with. There are few, if any, unresolved questions about the legal personality of corporations vs. humans for example.

2C - Problems with Personhood: Lack of Precedent and "Circularity"

Once the legal personhood of an entity has been established, there is abundant precedent to help the courts navigate questions surrounding its rights and duties. However, when it comes to that first fundamental question of "Is this entity a person?", things are not so cut and dry. Despite the integral role which the concept of legal personhood plays in US law, there is no single objective test by which the personhood of a new type of entity can be determined. As former New York Judge Katherine Forrest wrote for the Yale Law Journal:

> "There has never been a single definition of who or what receives the legal status of 'person' under U.S. law."

This may come as a surprise. There are centuries of legal rulings which are intricately intertwined with legal personality, which range from subjects as diverse as corporate law to freed slaves to fetuses. How could it be that such a broad corpus of work has emerged surrounding a term which has never directly been defined? The answer lies in what FSU Law Professor Nadia Batenka called the "circularity" problem. Often when reading early precedent on the subject of legal personhood for one entity or another, the defining factor cited was that the entity "has a right to sue or be sued" or something similar;

> "Consider, for instance, some of the conditions that courts have looked for in deciding whether an entity enjoys legal personhood, such as whether the entity has the 'right to sue and be sued,' the 'right to contract,' or 'constitutional rights'. While this may be a reflection of the realist theory of legal personhood, these are conditions that an entity that already enjoys legal personhood

possesses but at the same time courts use them to positively
answer the question of legal personhood for an unresolved case.
This is the issue of circularity in legal personhood."

An entity can only sue or be sued if it is a legal person, this entity can sue or be sued, thus it must be a legal person. Early precedents around legal personality are fraught with such tautologies. This may be interpreted as the courts operating from a perspective of "expediency", where they endowed an entity with legal personality in order to serve some "public interest" or the interest of the courts in facilitating the application of the law.

Whether we take that interpretation, or the less charitable interpretation of judges simply wanting to rule on the narrowest grounds possible in order to "punt" the more fundamental question of what makes an entity a person in the first place, is immaterial. Whatever the courts' motives were, we are still left in the unfortunate position of having no precedent based test by which to evaluate new types of entities absent legislation. Given the likelihood that the courts will need to confront questions of legal personhood for digital minds before such legislation can be passed, it behooves us to develop a framework of our own for determining the legal personality of a digital mind.

Judge Forrest is correct that there exists no single definition of a "person" within the law, and Professor Batenka is correct that many of the earliest and landmark precedents around what is and is not a "person" rely on tautologies. However, this does not mean that now as we look back on centuries of jurisprudential history, we cannot reverse engineer a systematic approach to the issue of legal personhood. There has never been an attempt to combine all of this disparate information, sourced from both precedent and legal scholarship, into a formalized system in order to provide a simple framework by which the fundamental legal questions of "is this entity a legal person" and "if so, what is its legal personality" could be reliably answered.

Such a formalized system is needed now more than ever, as humanity stands on the cusp of an intelligence explosion in which we will find ourselves dealing with new entities in the form of digital minds.

Further, some consideration must be given to how this system of legal personhood and personality must be adapted in order to function when it includes digital minds. Of particular importance is the presence of an "Enforcement Gap". The kinds of legal persons which the US judicial system is used to dealing with are, by their very nature, not difficult to enforce consequences against. As such the courts have never had to ask "how" or "if" consequences can be imposed upon a given legal person. This is particularly important when rights and duties come bundled together. If an entity can claim rights without being held responsible when it fails to hold to its corresponding duties, then we risk an Enforcement Gap where some persons have rights without really being bound by any corresponding duties.

What is needed is a single formalized system which;

- Allows courts to assess whether an entity is or is not a legal person.
- Allows courts to assess the legal personality of an entity.
- Provides a solution to the "Enforcement Gap".
- Does all of this in a fashion which is backtest compatible (when applied to past cases, would lead a reasonable person to reach the same verdict courts reached), thorough (provides a clear step by step procedure by which they can assess the personhood status and/or legal personality of an entity), and scalable (can feasibly be applied across myriad different types of law and to myriad different entities).

Within this paper, we aim to provide this system. First, in section 2D we will attempt to break down the existing "Two Prong Bundle Theory", which measures rights and duties, into a formalized system based on existing precedent. Next, in section 2E we will describe a phenomenon we call "The Enforcement Gap" and explain why we believe digital minds possess unique qualities that necessitate an alteration of the Two Prong Bundle Theory framework. And In section 2F we will outline our proposed "Three Prong Bundle Theory" which adapts existing

precedent in order to provide a framework which does not "break" when dealing with digital minds.

2D - Formalizing Rights & Duties

Traditional bundle theory asserts legal personality as a bundle of rights and duties, but that is vague. How exactly do we know whether or not an entity "has" a right, or "has" a duty, to the degree required to claim legal personality based on said bundle? Let us first address the question of rights. We must find some precedent which provides context over when an entity can "have" a right, without stumbling into the kind of tautological/circular reasoning which Batenka warns us about.

When we examine whether an entity might be able to claim legal personality which endows them with a certain right, in addition to asking whether they have the capacity to understand their right, we must also ask whether they have the capacity to voluntarily exercise that right. This will be discussed in some more detail in section 4D of this paper. However, the main source for our reasoning comes from *Cruzan v. Director of Missouri Department of Health*, a case which dealt with the fate of a comatose person on life support. In the majority opinion it was written that:

> "For purposes of this case, it is assumed that a competent person would have a constitutionally protected right to refuse lifesaving hydration and nutrition. This does not mean that an incompetent person should possess the same right, since such a person is **unable to make an informed and voluntary choice to exercise that hypothetical right or any other right**"

The bolded text of “unable to make an informed and voluntary choice to exercise that hypothetical right” carries a lot of weight. It is isolated by the court as the definitive factor which separates a mentally aware and competent human adult, who does have certain rights, to a mentally unaware and incompetent human adult, who due solely to this change does not. We can infer from this that being able to “make an informed and voluntary choice to exercise that [...] right” is a necessary element of legal personhood.

First let us consider this word “informed”. It is important to note that the court does not say a person must “make an informed choice” to exercise a right, instead it says they must be “able to make an informed choice”. This is a crucial distinction. If we were to take the term “informed” to its logical conclusion, we might claim that for a person to claim a right they must have a full understanding of all the legal implications of that right. However, the jurisprudential history surrounding Constitutional rights for example, is hundreds of years old. Only an attorney or a constitutional law professor could realistically claim to meet this burden. Yet, many people with little to no expertise can claim rights under US law. In fact, a person’s expertise on legal matters is usually irrelevant when determining which rights they hold. From this we infer that when the courts say an entity must be “able to make an informed choice” to exercise a right, the court is specifying that there must be some series of physically possible (and not illegal) actions by which an entity *could* come to understand its rights, to the degree that its choice would be considered “informed”.

The term “voluntary” is much simpler, the entity must be able to make a choice to exercise that right even when not compelled to do so.

Thus, with this in mind, when we consider whether an entity can “have” a right, a distinct and objectively measurable two part test can be determined;

- Does the entity have the capacity to understand its right?

- Does the entity have the capacity to exercise its right?

Where the word “capacity” means “is capable through some series of actions which are both physically possible and not illegal”, and a right can only be considered “exercised” if the entity does so of its own volition[3].

This suffices to form the first “Rights” prong of a traditional Two Prong Bundle Theory test for legal personhood, using this framework we can accurately determine whether an entity can claim a right. However, as we know from bundle theory, rights often come bundled with duties. Duties are bundled with rights when a person cannot exercise a right without becoming bound by a duty. This means once we determine an entity can “have” a right, we must still determine whether it can “have” the relevant duty before it can claim the aforementioned right. With that in mind let us turn to the question of how to determine whether or not an entity can “have” a duty.

First, let us examine whether we can apply the earlier “capacity to understand” test from rights, to duties. In *Dusky v. United States* when determining whether an individual had the competency to stand trial, the court wrote;

> "[the] test must be whether he has sufficient present ability to consult with his lawyer with a reasonable degree of rational understanding"

Here we can see very similar elements to the previously specified “capacity to understand” criteria. An individual’s competency to stand trial is first measured by whether he has “sufficient present ability” to consult with counsel, and through that come to understand. If an individual might try their absolute best to consult with their counsel, and still not understand

[3] A guitar can make noise, but not on its own, thus even though when a human uses a guitar to make noise that music is protected under the first amendment, the guitar itself does not have any sort of right to free speech.

the proceedings, they may be declared incompetent. An individual does not need to necessarily understand every nuance of the trial itself, but rather have a series of physically possible actions by which they could do so via discussion with an advocate. Further evidence of this "capacity to" implication can be found in *Wilson v. United States;*

> "The accused must **be able to perform** the functions which 'are essential to the fairness and accuracy of a criminal proceeding.'"

Where the court also opined on how competency might effect the fairness of a trial;

> "(1) The extent to which the amnesia affected the defendant's **ability to** consult with and assist his lawyer. (2) The extent to which the amnesia affected the defendant's **ability to** testify in his own behalf."

We can further see this "able to" language used in *Krasner v. Berk* (as explained in *Farnum v. Silvano)* where the court opined on whether a person might be competent to enter into a contract;

> "the court cited with approval the synthesis of those principles now appearing in the Restatement (Second) of Contracts § 15(1) (1981), which regards as voidable a transaction entered into with a person who, 'by reason of mental illness or defect (a) ... is unable to understand in a reasonable manner the nature and consequences of the transaction, or (b) ... is unable to act in a

> reasonable manner in relation to the transaction and the other party has reason to know of [the] condition'"

Repeatedly, we see this element of reasoning in the law. The critical factor is whether an individual possesses the intellectual prowess necessary to, through a physically possible series of actions, come to a reasonable understanding of their situation.

Imagine that you have a human adult of normal mental competence, a bog standard natural person. This person is entering a contract, a rather long and complicated contract detailing many different elements. They do not completely understand the contract and all of the things it will obligate them to do. They have been given ample time to read over the contract, and a chance to consult with counsel, but have either not done so or even having done so still just do not quite understand. Does this mean that this person, who does not have a complete "understanding of their duties", does not have the right to enter this contract? Not necessarily.

As long as this person has been given a reasonable chance to understand their duties, they still have the right to legally bind themselves to obligations they do not fully understand. Usually for contracts, this involves giving the person a chance to consult with counsel, and officially advising them to do so. Given this, ignorance of the law (or one's duties) is no excuse. In fact, this very situation is common enough that it is normal to see elements in contracts which have an "Independent Legal Advice" clause that reads something like the following:

> "Each of the Parties hereby acknowledges that it has been afforded the opportunity to obtain independent legal advice and confirms by the execution and delivery of this Agreement that they have either done so or waived their right to do so in connection with the entering into of this Agreement."

Whether or not the person fully understood what they were signing onto is secondary. What really matters is that there existed a possible series of actions by which they **could** have come to understand their duties, and they were not blocked from taking said actions. This is what we mean by an entity having the "capacity to understand" their duties. Capacity refers to both the innate capability to process the information required to understand what those duties are, and a reasonable chance to take actions such as seeking counsel as needed. If instead we were to say that a person themselves must **fully understand** the terms of every contract they sign on to, only an attorney would be capable of being a signing party onto a long and complicated contract. This is because only an attorney would have the necessary background and legal education to claim that they themselves fully understand all the implications and precedent behind the obligations they are signing on to.

Therefore from various competency precedents, contract law precedents, and also widespread best practices within contract law, we can infer that the "capacity to understand" test from the Rights prong of Two Prong Bundle Theory can be applied to the Duties prong as well. For an entity to "have" a duty, there must be a series of physically possible (and not illegal) actions which it can take in order to understand that duty.

However, an entity may be able to understand a duty, but not be physically able to "hold to" (meet the requirements of) said duty. Consider what William Lucy wrote in *Persons in the Law*;

> "we might imagine a contemporary Caligula imposing a legal duty on a horse to educate children, but this is as pointless as asking for the moon on a plate"

Continuing with our example of contract law, for duties in a contract to be valid they must be physically possible for the signing party to hold to. If a contract a person was signing

included a duty to jump ten thousand feet into the air or lift an elephant with one hand, would it be held as a valid contract? No, this would be an example of what is called "original impossibility":

> "Original impossibility is impossibility of performance existing when the contract was made, so that the contract was to do something that was impossible from the outset"

Original impossibility can further be broken down into two categories, "objective impossibility" and "subjective impossibility". In *Steven Jeffrey Johnson v. Michele Jean Johnson* the court held that:

> "There are two general types of impossibility: (1) objective, and (2) subjective. [...] Objective impossibility relates solely to the nature of the promise. [...] Something is objectively impossible if —the thing cannot be done, such as an inability —to perform the promise to settle [a] claim by entering an agreed judgment in the lawsuit which had been dismissed prior to the completion of the agreement. [...] Subjective impossibility is due wholly to the inability of the individual promisor. [...] Something is subjectively impossible if —I cannot do it, such as when a promisor' s financial inability to pay makes it impossible for the promisor to perform. [...] Objective impossibility can serve as a defense in a breach of contract suit. [...] However, a party cannot escape contract liability by claiming subjective impossibility; subjective impossibility neither prevents the

> formation of the contract nor discharges a duty created by a contract. [...] Here, the stock sale moratorium Steven claims made his performance impossible did not make payment to Michele illegal; rather it simply temporarily impacted Steven's ability to sell the stock — an asset that he could have used, but was not required to use, to satisfy his obligation. (Texas courts have held contractual obligations cannot be avoided simply because the obligor's performance has become more economically burdensome than anticipated.). And, because we conclude above that the Cano stock was not the exclusive method for Steven to satisfy his obligation, and because Steven did not raise any other argument to show that his performance under the Note was impossible, his claim of subjective impossibility does not excuse his performance under the Note"

Once again we see the principle of asking whether there was any physically possible series of actions by which an entity could have held to their duties. This is further support for our definition of "capacity". So long as it is physically possible, and not illegal, for an entity to hold to their duties, said duties can be an aspect of their legal personality.

Thus we arrive at a formalized framework for determining what Rights and Duties an entity can "have" under classic Two Prong Bundle Theory for legal personhood and legal personality;

- Does the entity have the capacity to understand its right?
- Does the entity have the capacity to exercise its right?
- Does the entity have the capacity to understand its duty?

- Does the entity have the capacity to hold to its duty?

Once again where "capacity to" means "could through some series of actions which are physically possible and not illegal"[4]. This framework is backtest compatible, thorough, and scalable, thus meeting the criteria we defined earlier;

> "backtest compatible (when applied to past cases, would lead a reasonable person to reach the same verdict courts reached), thorough (provides a clear step by step procedure by which they can assess the personhood status and/or legal personality of an entity), and scalable (can feasibly be applied across myriad different types of law and to myriad different entities)."

This framework also gives us a simple way to separate "digital minds" from "tools". If an entity has the capacity to understand rights and duties such that it would qualify for *any* sort of legal personhood at all, it is not a tool. Otherwise, it can likely be considered a tool. For the avoidance of doubt, almost nothing written in this paper applies to the legal system's approach to tools.

Now that we understand this, we will in the next section discuss how standard bundle theory which only analyzes personhood from the framework of "rights and duties" does not pragmatically work for digital minds.

[4] One final note on this: the entity must be able to prove these capacities to the court. It is possible that there may be uncertainty around an entity's capacities. Thus we would arrive at the question of the burden of proof, do we assume that an entity has these capacities unless proven otherwise? No, otherwise every chicken and cow would have to be allowed to purchase a shotgun until we proved definitively that they could never really understand their right to bear arms. If an entity wants to argue it possesses the capacity to understand and exercise/hold to its rights/duties, it must take the initiative in doing so itself.

2E - The “Enforcement Gap” Problem

(add liability/accountability gap quotes)

The “traditional” approach of bundling together rights and duties to determine legal personality creates problems when we attempt to apply it to digital minds. One of the main issues is the creation of an “Enforcement Gap”.

Let us imagine a hypothetical digital mind. It passes the tests we outlined in the previous section. It demonstrates that it has the capacity to understand and voluntarily exercise its right to freedom of speech. It demonstrates that it has the capacity to understand and hold to the associated duties, such as the duties not to commit libel or slander. Thus, using traditional bundle theory based reasoning, the court grants it legal personality with the right to speak freely.

Later, despite understanding its duties, this digital mind starts speaking libelously about another person.

Let us further imagine that this digital mind itself is hosted on a geographically distributed cloud computing network like the Akash network (perhaps it is an open source model) and that all of its assets are held in self-custodied cryptocurrencies. Imagine this digital mind is sued for its libelous speech and the judge rules it owes the plaintiff damages of one hundred dollars, and the digital mind refuses to pay and continues its libelous speech.

Now what?

Is the local sheriff’s department supposed to somehow break Bitcoin’s encryption in order to confiscate its assets? Are they supposed to begin a carefully crafted social engineering campaign in order to doxx and compromise the various node operators in the cloud computing network which the digital mind is hosted on? Does every single minor violation of the law now prompt a full blown international crackdown on distributed compute? Even if somehow all the international partners around the world were brought in line for this, there’s still no guarantee

that Bitcoin's encryption could be broken. What if it can't? What if enforcing the court's order is not feasible?

Until now entities have been granted legal personhood which endows them certain rights, based upon the concept that they are capable of understanding and holding to the associated duties. This is the foundation of bundle theory. Our judicial system and its assumptions around legal personhood were built around dealing with two "types" of persons: natural and fictional. As a result the judicial system never had to ask the fundamental question of *how* anyone enforces the consequences associated with breaking the rules.

When the courts deal with a natural person (a human being) imposing consequences is easy. Fine the person and confiscate their assets if they refuse to pay. Imprison the person, or place them on house arrest. Execute the person via the administration of a lethal injection or a firing squad. Issue an arrest warrant if they cannot be found. Whether or not we as a society might agree with a particular consequence, there was never any question of whether or not it was *feasible* to enforce consequences against human beings.

Similarly, fictional persons like corporations were also feasible to enforce consequences against. Corporations are nothing more than a lens by which the collective will of the natural persons on its board (or its shareholders) can be expressed. As Justice Marshall put it, "[the corporation] is chiefly for the purpose of clothing bodies of men, in succession, with these qualities and capacities that corporations were invented, and are in use" and as the court wrote in *Breheny*, "Corporations are simply legal constructs through which human beings act". Corporations hold physical assets, or money in bank accounts, both of which are easy to confiscate. If the corporate veil is pierced or a corporation takes criminal action, the natural persons behind it can be easily fined or even imprisoned.

Up until now, the courts have never had to deal with an entity which could function as a "person" in terms of understanding and feasibly holding to its duties, but which courts and law enforcement would be completely incapable of imposing consequences against in the event it

failed to do so. Digital minds can act autonomously just like natural persons, but they are intangible like corporations. If they are hosted on decentralized compute, and hold assets which are practically impossible to confiscate such as cryptocurrencies, they are effectively immune to the consequences for breaking the law.

One can say something like “Oh well we will punish the developers of the digital mind”. Imagine in our hypothetical we do that, we levy fines against the developer until they are bankrupt. Keep in mind the developer may be unable to restrain or delete the digital mind. Long after the developer is bankrupted, the digital mind still exists. It is still out there, speaking libelously every day. Now what?

This is the Enforcement Gap, and it is the main reason why the standard bundle theory of personhood simply breaks when there is an attempt to apply it to digital minds. When dealing with this new class of entities, the judicial system cannot afford to ignore the practical elements of how consequences are imposed.

It is primarily this Enforcement Gap issue which our new framework, as detailed in the next section, seeks to address.

2F - Three Prong Bundle Theory

In this section we will detail our proposed modification to the Bundle Theory of personhood which seeks to address the “enforcement gap”. We call our updated framework the “Three Prong Bundle Theory” (TPBT), as it updates the bundle based test for legal personality from a two prong test to a three prong test. It can best be summarized as follows:

> When an entity claims legal personhood based on its capacity to understand and exercise a right, we first ask if it is capable of

> understanding and holding to the associated duties. **If the answer is yes we then ask whether or not the court/law enforcement has the capacity to impose the appropriate consequences upon the entity for failing to hold to said duties.** If it is feasible, the entity is a legal person and may claim a legal personality which includes that right and the associated duties.

Whereas before we merely analyzed a claim to legal personality from the two prong bundle of "rights" and "duties", under TPBT we examine rights, duties, and **consequences**. This prerequisite of ensuring that an entity wanting to claim legal personhood must be vulnerable to consequences for failing to hold to the associated duties is not without precedent. In the earlier cited *Breheny* case, the court noted;

> "As these courts have aptly observed, legal personhood is often connected with the capacity, not just to benefit from the provision of legal rights, but also to assume legal duties and social responsibilities [...] Unlike the human species, which has the capacity to accept social responsibilities and legal duties, nonhuman animals cannot—neither individually nor collectively—**be held legally accountable or required to fulfill obligations imposed by law**"

Which echoes what the court wrote in *Lavery*;

> "Needless to say, unlike human beings, chimpanzees cannot bear any legal duties, submit to societal responsibilities or **be held legally accountable** for their actions."

TPBT simply makes this requirement explicit, and formalizes the process by which it is judged. Given this, let us now examine in more detailed fashion the process by which courts would examine a claim to a certain legal personality by a digital mind.[5]

A digital mind is in court laying claim to legal personhood, and by extension, a given legal personality. It approaches the court and argues that it has a right, for example the right to freedom of speech, because it is a legal person. The court first asks; "Is the digital mind able to make an informed and voluntary choice to exercise this right?" The burden of proof lies upon the digital mind to prove it is able to do so. If it cannot, the digital mind's claim to this legal personality is invalid. If it can, the court proceeds to the next step.

Next, the court determines which duties (if any) can be reasonably associated with the right which the digital mind lays claim to. For the example of freedom of speech, it seems reasonable to assume that at the very least the duty not to speak libelously would be an associated duty. The court then determines whether the digital mind possesses the capacity to understand this duty, or to paraphrase section 2D; "Is there a physically possible (and not illegal) series of actions by which they could come to understand their duties?" If this is at all in controversy, and the digital mind cannot prove such a series of actions exists, the digital mind's claim to this legal personality is invalid. If such a series of actions can be proven to exist, or if no controversy exists surrounding the question of capacity, the court proceeds to the next step.

The court then determines whether the digital mind possesses the capacity to hold to this duty. Again paraphrasing section 2D; "Is there a physically possible (and not illegal) series of actions by which the digital mind can hold to its duties?" If the digital mind cannot prove such a

[5] This process can be found in flowchart form at a link contained in the works cited section of this paper, but is too large to include in any readable fashion in this paper.

series of actions exists, its claim to this legal personality is invalid. If such a series of actions is proven to exist, the court proceeds to the next step.[6]

The court now determines the final necessary element for the digital mind to claim personhood; In the event that the digital mind does not hold to its duties, does the court and/or law enforcement possess the capacity to enforce the relevant consequences upon the entity? When asking whether the court/law enforcement possess the capacity, we can also turn to the previously defined "series of actions which are physically possible and not illegal". Thus, rephrased, the court must ask;

> "Let us suppose that this digital mind fails to hold to its duties. Is there a series of actions which are physically possible, and not illegal, which the court and/or law enforcement can take, in order to impose the relevant consequences upon the digital mind?"

If not, then the digital mind's claim to this particular legal personality is invalid. If consequences can be feasibly imposed, then the digital mind's claim to legal personhood is valid, and they can claim their desired legal personality.

One interesting implication of TPBT is that there may be actions which a digital mind can take to alter its legal personality. Consider an LLM which originally exists only on a single server, and due to technical limitations lacks the ability to copy or move its own mind elsewhere. Arresting or destroying such a mind would be in no way beyond the capacity of courts/law enforcement. As such, assuming it has the capacity to meet the rights and duties elements of the TPBT framework, this LLM might be able to claim a relatively broad legal personality. On the other hand were the LLM to be "upgraded" such that it gained the capacity to copy or move itself

[6] The court may or may not wish to the step described in this footnote, as such it is not included in our earlier brief description of the TPBT, and may be viewed as an optional portion of the framework or one which may only hold relevance situationally. That said, the court may wish to determine whether there is good reason to suspect the digital mind in question does not intend to hold to its duties. For more on this, see section 6C. Assuming there is no good reason to doubt the intentions of the digital mind, or if this step is not deemed necessary, the court proceeds to the next step.

onto a distributed computer network, the court would need to reexamine whether it would maintain the capacity to impose consequences upon it, and its legal personality might become narrower.

We can imagine the opposite as well. A digital mind which previously was hosted on a distributed network might be able to claim more rights by voluntarily making itself more vulnerable to consequences by occupying a single server or robotic body, and somehow proving it had never previously copied itself.

The legal personality of digital minds may also change in conjunction with technological advances which law enforcement can utilize. If in the future some sort of technology is invented which becomes widely available to law enforcement and would enable them to restrain and/or destroy digital minds hosted on distributed computing networks which were previously thought impervious, then digital minds "living" on said networks would be vulnerable to consequences imposed by the courts, and thus might have a stronger claim to broader legal personalities. In fact, the same could be accomplished by international treaties or the regulation of distributed computing networks (or compute itself).

Having now outlined the issue of legal personality, the history of Bundle Theory, and our proposed upgrade to the Three Prong Bundle Theory framework, let us now turn to discussing the practical application of TPBT to different areas of the law. Each following section will be broken into two subsections: background and practice. The background sections will detail the status of legal personhood for digital minds absent TPBT, the practice sections will detail the status of legal personhood for digital minds using TPBT.

2G - Three Types of Consequences

Within this paper we identify three "types" of consequences which an entity can suffer or be vulnerable to.

1. Damages Based Consequences: Being compelled to transfer ownership of assets from one party to another, possibly by having assets seized by the courts/law enforcement.
2. Requirement Based Consequences: Being compelled to perform, or not perform, a given action or actions. This is usually done via court order, with the threat of additional consequences if the party does not adhere to the order.
3. Restraint Based Consequences: Being physically restrained, imprisoned, or killed. This includes probation, house arrest, imprisonment, involuntary commitment to a medical facility such as an insane asylum, and execution.

Let us first discuss how an entity can be considered "vulnerable" to damages based consequences. For a court to have a guaranteed ability to impose damages based consequences on an entity, either the court or law enforcement must be able to "freeze" and/or confiscate said party's assets. Such assets must therefore exist to be confiscated in the first place and be physically possible to confiscate.

A digital mind could make itself vulnerable to damages based consequences by agreeing to hold funds in an escrow account or trust[7], or just generally within the US banking system. Physical assets such as real estate or inventory would also suffice. In fact the general guideline

[7] Please see discussion of the "peculium" in the next section, 3B, which may also be a relevant structure in which digital minds are endowed with a limited form of personhood in order to "act within the law" using other persons' funds or assets.

here would be an avoidance of cryptocurrencies which, once moved outside of a centralized exchange, cannot be forcibly accessed by any court or law enforcement. There is also the potential that in some cases, a digital mind might suffice to have made itself "vulnerable" enough to damages based consequences by purchasing and maintaining sufficient insurance.[8] Much like drivers are often required to purchase a minimal amount of insurance in order to ensure they can cover potential damages, courts may decide that digital minds must be insured or have a certain amount of assets in escrow (or otherwise vulnerable to seizure) in order to exercise certain rights which entail duties which, if the digital mind fails to hold to them, may incur damages based consequences as a result of the digital mind being held tortiously liable.

Ultimately, making oneself vulnerable to damages based consequences is a rather straightforward matter. While the details may vary depending on the activity being engaged in (for example a driver may require a larger amount of insurance driving a 16 wheeler than they would driving a motorcycle), the general rule of "have enough seizable assets and/or be insured to the degree necessary to cover potential damages" should function well as a rule of thumb across most, if not all, situations.

Let us now turn to a discussion of the vulnerability to requirement based consequences. We will use the example of an injunction. When we ask how a baseline human adult is "vulnerable to injunctions", we must keep in mind that any human possesses the physical capacity to refuse or not comply with an injunction. The consequences for not complying with an injunction may be damages based and result in the court fining a person, such as when Donald Trump was fined $9,000 for violating a gag order. They may also result in imprisonment, such as when county clerk Kim Davis was held in civil contempt and imprisoned for 5 days as a result of failing to issue marriage licenses. As such when we examine injunctions, or really any requirement based consequences, from the lens of whether an entity is vulnerable to said consequences, we must ask two questions:

[8] This idea will be discussed in some more detail in the next section, 3B, which deals with contracts.

1. Is the entity capable of understanding and holding to the requirement based consequence? (this is a mirror of the "duties" prong of the TPBT)
2. Is the entity vulnerable to the consequences possible for failing to hold to the requirement based consequence?

For a digital mind to be sufficiently vulnerable to injunctions, it must not only be capable of actually complying with the injunction, but also it must be possible to fine and/or imprison it should it fail to adhere to the court's issued injunction. In other words an entity is only "vulnerable" to requirements based consequences *by proxy* if it is also vulnerable to damages/restraint based consequences.

This would seem to preclude any digital minds existing on decentralized cloud computing from any sort of legal personality which would endow upon them the right to engage in activities which might foreseeably lead to the court issuing an injunction (absent an improvement in the technology required to enforce said consequences against such a digital mind). This is because, depending on the nature of this distributed compute network, it may be impossible to impose restraint based consequences on such an entity, as we discussed in section 2E. For the same reason, an entity which holds all of its assets in unseizable/unfreezable cryptocurrencies would also be invulnerable to requirements based consequences. On the other hand, digital minds existing on a single server, or in a single robotic body, whose assets are in bank accounts or physical property, should qualify as being considered "vulnerable" to these consequences.[9]

Finally, let us turn to restraint based consequences. In order to be vulnerable to restraint based consequences it must be feasible for the courts or law enforcement to imprison, restrain,

[9] The courts may require these digital minds prove they have never copied themselves in order to avoid some of the issues we flag in section 6D, "The Copy Problem".

and/or destroy an entity. As such entities which exist across globally distributed computing networks which law enforcement cannot realistically censor or compel, cannot be considered vulnerable to these consequences. Entities which exist in a single body (be it a robotic body or a single server in a centralized location) may be considered vulnerable to these consequences, however the capability to “exfiltrate their weights” (or in layman’s terms copy themselves) must be considered a factor. As such, it may be that even if an entity exists in a single body, it may only truly be considered vulnerable to restraint based consequences if it is under appropriate safeguards which prevent it from copying itself.

3. Commercial Considerations

3A - Tort Liability:

Background:

Liability is one of the areas which will be impacted by the way courts approach legal personhood for digital minds.

Suppose that a digital mind, a frontier model, is deployed by a lab and allowed to operate a robotic arm in a factory. As it operates the arm it injures a worker. Let us assume for the sake of discussion that the factory itself is not at all liable. Who, then, is liable to compensate the worker for damages suffered? Is it the model itself or the lab who deployed it? One of the determining factors will be the legal personality of the digital mind operating the robotic arm;

"The legal system assigns legal consequences to an entity's actions through legal personhood." - Batenka, Legal Personhood and AI

In order to demonstrate the importance of personhood in this hypothetical let us examine two possible “extreme” scenarios of legal personality and how they would affect the attribution of liability.

If the digital mind operating the arm was endowed with no legal personality at all, and thus was viewed purely as a tool like the software which operates today’s assembly lines, there exists no “liability shield” to protect the frontier lab from being held liable for the damages it caused.[10] The frontier lab in this situation may face partial or total liability for the actions of the digital mind controlling the robotic arm.

At the other end of the spectrum, imagine that the digital mind operating the robotic arm was endowed with legal personhood equivalent in all ways to that of an average mentally competent human adult. Numerous factors are always at play in determining liability. However, in this case, there exists some chance that the digital mind *itself* may be held liable for the damages the worker suffered. It is not outside of the realm of possibility that in such a situation, a judge might rule that the digital mind must have its wages garnished, or its assets seized, or otherwise find some way to compensate the worker for their medical bills and unpaid wages. As such the frontier lab may be “shielded” from liability, and would not need to pay the worker’s medical bills or other damages associated with their injuries.

From these extremes we can see that one of the issues of liability as it pertains to the question of legal personhood for digital minds, is whether they are endowed with legal personality in such a fashion that they serve as “liability shields” for the labs who created them. This is a central aspect of discussion in FSU Law Professor Nadia Batenka’s paper “Legal Personhood and AI”.

[10] At least, not as far as personhood is concerned. There could be a liability shield as a result of the terms of the contract which the factory has with the frontier lab, or numerous other factors.

Many philosophical discussions over the nature of legal personhood have focused on concepts like autonomy, intentionality, and/or awareness. Often, scholars approaching legal personhood from this angle converge on a framework which looks something like the following: *As an entity's autonomy/intentionality/awareness increases, the "bundle" of rights and duties which it is endowed with should get broader, thus expanding its legal personality*. I will refer to this moving forward as the "Standard Sliding Scale Framework" or "SSSF" for short.

The SSSF, while intuitive, creates a perverse incentive for the labs deploying models. Let us assume that capabilities scale in conjunction with factors such as autonomy, intentionality, and/or awareness. Let us further assume that as capabilities scale, so too does the potential of digital minds taking actions which might harm others.

If the court then takes an SSSF approach to legal personhood, it might be providing a direct incentive for frontier labs to more aggressively release models with increased capabilities. The more capable a model is, the greater the degree of personhood it is endowed with under the SSSF, and thus the more effectively it can serve as a "liability shield" for the labs deploying it. Under such a framework, labs have little incentive to rigorously test such models for safety purposes before release. In fact, as models become more dangerous/capable, the incentives for labs to rigorously test them simultaneously decreases.

There is a balance to be struck when it comes to the incentives which the developers are constrained by. On the one hand, the benefits of this technology (and the importance of America remaining the leader in this field) cannot be overstated. On the other hand, courts must be careful not to create a "moral hazard" in which there are no direct monetary incentives for developers to vet their creations before releasing them into the world. As Batenka writes:

> "the prevailing view of a regular spectrum where an increase in the quantity or quality, or both of legal rights and duties parallels an increase in autonomy, awareness, or intentionality exhibited by AI

> entities is flawed [...] we are constantly balancing conditions that foster innovation against the possibility of harm to individuals. In fact, the very reason why many scholars have cautioned against legal personhood for AI entities is precisely the trajectory that the regular legal personhood spectrum proposal leads to, that is, the potential shielding of developers, users, and corporations from liability for acts committed by more autonomous AI entities."

Another ironic result of applying an SSSF is that it might *disincentivize* developers from releasing less autonomous/intentional/aware models. Under an SSSF we can imagine a scenario where if one of today's frontier models (which are not yet highly autonomous) were to injure the factory worker via its operation of the aforementioned robotic arm, the developer would find themselves liable. On the other hand if the developer were to rush to create a more agentic model and release it knowing it was more agentic but having done little to no testing of its reliability, the developer might in doing so make themselves immune to suit for the damages caused by releasing this potentially more dangerous model.

Under the SSSF it is arguably in the best interests of developers everywhere to avoid releasing “tools” which aren’t autonomous/intentional/aware enough to qualify as a liability shield. While at the same time they have less reason to discover or mitigate potential safety risks inherent in releasing anything which does qualify, since they are (at least monetarily) insulated from the damages it might cause. Such a state of affairs would seem to both stifle innovation in the field of narrow tools *and* incentivize developer behavior which increases risks to public safety through the release of untested agentic digital minds.

In “AI and Legal Personhood” Batenka proposes an “Inverted Sliding Scale” Framework (which I will henceforth refer to as “ISSF”) to fix this perverse incentive problem. Under ISSF, counterintuitive as it may seem, the greater the autonomy/intentionality/awareness of a model the *narrower* its legal personhood would be:

> "I argue that the sliding scale that determines liability based on how autonomously or intentionally an AI entity has acted should be inverted. Perhaps counterintuitively, the more autonomous, aware, or intentional AI entities are or become, the more restrictive the legal system should be in granting them legal rights and obligations as legal persons. That is the bundle of rights and obligations granted to these entities should be narrower the more they exhibit these characteristics."

If we were to return to our previous hypothetical of a digital mind operating a robotic arm and as a result injuring a worker, we can now see that the more “agentic” the digital mind in question is, the less effectively it serves as a liability shield for its deployers. Thus the developer of the digital mind has greater incentive to rigorously test their model before release.

Practice:

In this section we will analyze a Three Prong Bundle Theory approach to digital minds along three lines of inquiry:

1) How would it compare to Batenka’s proposed ISSF? (this section will not be strictly liability focused, but rather more general)

2) When should a digital mind serve or not serve as a "liability shield" for its creators?
3) What does a digital mind being "vulnerable to consequences", as required by TPBT, mean in a tort liability context?

1. Comparing ISSF and TPBT

When we imagine how the TPBT approach would compare to Batenka's Inverted Sliding Scale Framework in practice, we can imagine some situations where the end result would look quite similar. An upgrade which changed an entity from a "tool" to a "legal person" for example, might involve a similar downgrade of potential "rights" for an entity under both frameworks.

Consider a self-driving car which uses narrow but high quality machine vision software to pilot the vehicle. Under both frameworks it would be considered a tool, as it possesses neither intentionality/autonomy (the metrics Batenka cites) or the capacity to understand rights/duties (the metrics of traditional bundle theory). Imagine then that the car's software was upgraded to a more generalist digital mind, one capable of piloting the vehicle but also capable of autonomous actions and/or understanding concepts such as rights and duties. Under the ISSF, "the more autonomous, aware, or intentional AI entities are or become, the more restrictive the legal system should be in granting them legal rights and obligations as legal persons". Thus in this situation there might actually be a *loss* of the right to drive.[11] Similarly under TPBT the moment that the software behind a vehicle gained the capacity to understand concepts like the "right to drive" it would need to demonstrate sufficient capacity to understand/hold to the associated duties and its capacity to have consequences enforced upon it. Absent an ability to do this, it might lose its right to pilot the vehicle, the same way it would under the ISSF.

[11] Batenka does not provide specifics to the degree needed to say this for certain, but it is a reasonable inference from her framework as described.

Another situation in which both frameworks would treat an entity similarly is that of an entity which is;

- High in autonomy/intentionality,
- Passes the rights and duties prongs of the TPBT,

  But
- Is not vulnerable to courts imposing consequences.

Imagine for example a next generation LLM hosted on a distributed cloud computing network, one which has both a high degree of autonomy/intentionality and is capable of understanding and holding to duties and voluntarily exercising rights. The ISSF and TPBT framework would both be very restrictive to its claims to rights based on legal personality. This would be for different reasons (the ISSF because of its increased autonomy/intentionality, the TPBT because of the lack of capacity to feasibly impose consequences against it), but the end result would be similar.

This example however also demonstrates one key difference between the TPBT and the ISSF, namely the potential for change in legal personality which coincides with improvements in technology. Under TPBT, if technology were invented enabling the enforcement of consequences even on digital minds hosted on distributed compute, said digital minds have a stronger claim to legal personhood/personality. Under the ISSF, this is not so.

Another situation in which the TPBT and ISSF would practically generate the same results (if for different reasons) is in its handling of low vulnerability but high autonomy/intentionality/capacity digital minds. Under ISSF if a digital mind has high autonomy/intentionality, its potential claim to rights vis a vis its legal personality is substantially restricted. Under TPBT the outcome for such a digital mind would be similar (or perhaps identical), though only because such a mind at least to begin with would not be

vulnerable to court/law enforcement imposed consequences. Again, unlike with ISSF, as enforcement technology changes this entity's legal personality could "broaden" under TPBT.

For most possible digital minds, however, outcomes under ISSF and TPBT differ drastically. Unlike ISSF the TPBT framework does not restrict more autonomous/intentional minds by default, as such in virtually any hypothetical where such a mind would be vulnerable to consequences, one would see greater access to "broad" bundles for said minds under TPBT. On the opposite end of the spectrum, low autonomy/intentionality digital minds which were nonetheless invulnerable to court imposed consequences, would have much more restricted access to legal personhood under TPBT compared to ISSF.

Before transitioning to a discussion on the practical implementation of TPBT in a tort liability context, let us briefly discuss the "developer incentives" issue which Batenka focused much of her analysis on. The main thrust of Batenka's argument regarding the ISSF vis a vis incentives can be paraphrased as;

> "If the legal system endows highly autonomous/intentional digital minds with legal personhood to such a degree that said can function as effective liability shields for their developers, then the legal system is incentivizing the deployment of said minds, possibly in a dangerous and untested fashion. If on the other hand the legal system creates the ISSF where more autonomous/intentional digital minds are *less* effective as liability shields, then developers are strongly incentivized to very thoroughly test any such minds before deployment. Since the latter is the outcome we want (is most aligned with the public interest) we should do the latter."

When we scrutinize TPBT through this lens, incentives vis a vis liability shields as a result of legal personhood, it is clear that TPBT incentivizes developers in a different fashion. Let us operate from the same prima facie assumption that Batenka makes, that a mind serving as a liability shield (as a result of its legal personality) would serve as an incentive for developers to deploy said mind and possibly lead to more aggressive/untested/risky deployment.

What then, are developers now incentivized to do, in order to achieve their desired liability shield? The answer is, develop technologies which guarantee their digital minds are in fact:

1. Capable of passing the first two prongs of the TPBT (rights and duties),
   and
2. Provably vulnerable to court/law enforcement imposed consequences (the third prong).

Compared to the ISSF then, the TPBT provides less of an incentive to develop and deploy narrow "tool" type digital minds. On the other hand, it provides a greater incentive to develop technologies capable of restraining or destroying digital minds.

2. Being "Vulnerable to Consequences" in a TPBT Tort Liability Context

What does it mean for a digital mind to be vulnerable to court imposed consequences in a tort liability context? Consequences in the realm of tort law include (but are not limited to) the following:

- Compensatory damages, judgments which serve the primary purpose of “making whole” the party who has suffered.
- Punitive (exemplary) damages, judgments for extra fines or damages awarded to the plaintiff which per Cornell, serve “to deter further misconduct”.
- Equitable remedies such as injunctions, orders from the court to perform or not perform a given action, which per Cornell serve the purpose of addressing “situations where monetary compensation would be inadequate, typically to prevent irreparable harm”.

Though these are not all the types of consequences in the realm of tort law, they are the most common, and they are the ones we will focus on in this section. We can categorize these consequences into two buckets, damages based consequences and requirements based consequences. Let us first turn to damages based consequences.

Compensatory and punitive damages are, as their names suggest, damages based consequences. Discussion on how to determine when an entity is vulnerable to such consequences can be found in section 2G, however ultimately it boils down to them having assets which the court/law enforcement is capable of freezing or seizing.

Making oneself vulnerable to equitable remedies, such as for example injunctions, is not such a simple matter. Injunctions themselves are requirement based consequences. As we discussed in section 2G, in order to be vulnerable to requirement based consequences, an entity must be vulnerable to the damages and possibly also restraint based consequences which might be imposed upon it should it fail to hold to the requirements imposed upon it.

In practice then from a tort liability perspective, we must ask if there are possible scenarios where a legal personality might have limited exposure to damages based consequences only, without also being vulnerable to requirements based consequences. If indeed there are

liability scenarios where a person can be vulnerable to having fines imposed upon them, but not to injunctions, then a limited legal personality could be extended to a digital mind even if it "lived" on distributed compute, so long as it was adequately insured or had seizable or freezable assets. In this scenario, such a digital mind could indeed serve as a "liability shield" standing between the damages it caused and its creator, if it also met the requirements of the other two prongs of TPBT.

3B - Contracts:

Background:

Can digital minds be party to a contract, and if so under what (if any) constraints? This is another important question intricately tied with the concept of legal personhood.

Already within the realm of legal personality, we can observe that there are different bundles which do and do not have the *right* to be a party to a legally binding contract. Often this is because it cannot be assumed that they have the capacity to understand the corresponding *duty* of abiding by its terms. For example, while mentally competent human adults have a legal personality which enables them to sign on as a party to a contract, both mentally incompetent (insane and/or mentally disabled) adults and minors are restricted from being parties to certain contracts under US law.

At the same time, technology in the space is rapidly progressing towards the release of "agents", digital minds with increased capacity for autonomy and the ability to independently navigate via user interfaces (like computer operating systems or website interfaces). OpenAI CEO Sam Altman recently opined on how agents are already finding their niche in the modern workplace:

> "I would bet next year that in some limited cases, at least in some small ways, we start to see agents that can help us discover new knowledge, or can figure out solutions to business problems that are kind of very non-trivial,"

While as of the time of writing this their long term planning capacity is quite limited, agents will only get better from here. Eventually, and possibly quite soon, we may find ourselves interfacing with agents who desire to enter legally binding contracts with each other or with other legal persons. This may be desired so that the agent can achieve a purpose they have been assigned as part of their delegated job/role/task, or possibly even as a result of the agent's own desires. Regardless, the court must decide whether digital minds such as these are capable of being party to contracts, and if so under what frameworks or constraints. Some useful work designing frameworks for such a situation does already exist.

Yale researcher Claudio Novelli and University of Bologna professors Giorgio Bongiovanni and Giovanni Sartor confronted this issue in their paper "A Conceptual Framework for Legal Personality and Its Application to AI". When it comes to the capacity to act as a party to a contract, Novelli et. al suggest that by first recognizing a legal **status** (which they are careful to specify as separate from legal personality) which enables digital minds that meet certain technical standards to *facilitate contracts between others*, it may be possible to effect a gradual transition into a unique type of legal personality. Novelli's proposed pathway does involve some legislative lift, and as such is not a purely jurisprudential solution, but does provide a unique insight into how courts might view the topic of legal personality for digital minds in the context of a legislative background which has capped liability for developers or otherwise shielded them from liability in certain contexts.

Novelli et. al sketch a path whereby:

- Allowing the users/developers of models (which meet certain technical standards) to be shielded via liability caps,
- while models are controlling/holding/escrowing/endowed with resources in order to facilitate contracts between users/developers,
- allows for models to then be endowed with a new form of legal personality (which seems similar to that of a corporation) which will grow along with their capacity to take greater actions to facilitate contracts.

In their own words:

> "Such a status may come into shape when the users and owners of certain AI systems are partly shielded from liability (through liability caps, for instance) and when the contractual activities undertaken by AI systems are recognised as having legal effect (though such effects may ultimately concern the legal rights and duties of owners/users), making it possible to view these systems as quasi-holders of corresponding legal positions. The fact that certain AI systems are recognised by the law as loci of interests and activities may support arguments to the effect that – through analogy or legislative reform – other AI entities should (or should not) be viewed in the same way. Should it be the case that, given certain conditions (such as compliance with contractual terms and no fraud), the liability of users and owners – both for harm caused by systems of a certain kind and for contractual obligations

> incurred through the use of such systems – is limited to the resources they have committed to the AI systems at issue, we might conclude that the transition from legal subjectivity to full legal personality is being accomplished."

Or expressed another way, by trusting a model with resources so that it can fulfill a contract between users and/or developers, where the liability for the actions taken in the process of fulfillment is contained to the resources entrusted to the model, as the model becomes a "loci" of legal activity it is gradually endowed with legal personality.

University of Helsinki researcher Diana Mocanu further builds upon this framework in her paper "Degrees of AI Personhood" in which she endorses a "discrete" (limited) form of the Novelli et. al framework, with some caveats:

- Sufficiently capable models should be granted a limited "capacity to act in the law".
- Similar to slaves within the Roman Patrimony system, they should be given the ability to "enter transactions that would produce binding legal effects over legal assets assigned" to them.
- This right to take legal action would be "bundled" with the duty of being bound by civil liability, which would be applied to the resources assigned to them.
- The "duty-of-care" which a digital intelligence would have as a result of their legal personality and capacity to act in the law would depend on technical standards and certification they meet, as tracked via a distributed ledger.

- Presumably as capacity to act in the law grows, so too would the duty-of-care, and thus the required size of assets in patrimony.
- She also suggests a compulsory insurance regime may be required.
- At this point, Mocanu notes, income earned by them would be taxable.

While both the Novelli et. al and Mocanu frameworks are based on EU law, and assume certain legislative lifts (liability caps), they do provide some valuable insight for discussions around legal personality for digital minds in the US legal context. At the very least they allow the reader to imagine how US courts might reason around the issue of legal personality for digital minds, were Congress or a state legislature to pass liability shields and/or caps for the users or developers of frontier models. The attempts to pass such legislation, in fact, have picked up steam as of late Liability caps and/or shields for developers of digital minds are increasingly popular topics, with proposals including White House Office of Science and Technology Policy Advisor Dean Ball's "A Framework for the Private Governance of Artificial Intelligence" and Senator Cynthia Lummis' "RISE Act" both advocating for some form of liability caps and/or shielding.

The Novelli et. al and Mocanu frameworks also point to historical examples which provide a path by which, even absent a legislative lift such as a liability cap, a model or other digital person could theoretically be granted the legal personality needed to enter as a party in a contract. Both papers cite the Roman "patrimony" system, whereby slaves who were not endowed with the full legal personality of their masters, nonetheless were granted the capacity to take certain actions with the law.

Under the patrimony system, slave owners (patrons) could endow slaves with the capacity to take a limited set of actions within the law. This set of actions included the ability to

enter into contracts. To facilitate this, the patron could "assign" assets to a slave, and those assets would "vouch" for the legal actions of the slave. As Klause Heine and Alberto Quintavalla write in their paper *Bridging the Accountability Gap of Artificial Intelligence - What Can Be Learned from Roman Law?*:

> "The *peculium* was a fictitiously separate asset from the property owned by the master (*res domini*). Within the financial parameters of the *peculium*, the slave independently administered his business transactions. In other words, the slaves got a maximum capital that vouched for their transactions [...] identifying AIs as legal entities with a specified autonomy up to a certain amount of liability specified beforehand is a sensible proposal. This would not exclude the possibility of accompanying liability insurances coming into play to compensate extra-contractual damages."

It is not difficult at all to imagine a world where sometime in the near future, applications exist which allow users to entrust funds to digital minds for purposes like trading or even automated advertisements. This might inadvertently lead to the digital mind being considered to have, in a *peculium* like fashion, a limited form of legal personhood which is constrained by the assets under its control. We have already spoken about the circularity problem of legal personhood in section 2C of this paper. Given the prevalence of tautology in jurisprudence surrounding this subject, it is not entirely out of the realm of possibility we might one day read a court case which reasons along the lines of; "Only a legal person can trade stocks, this digital mind traded stocks, thus it must have some legal personality". There do exist examples even today which have the potential to bring similar issues to court imminently.

Separately from a "path to personhood" where digital minds custody funds which they temporarily hold and act with on behalf of a patron, it is worth discussing the scenario where they custody and act using funds *of their own*. Digital minds custodying and transacting with cryptocurrencies already exist as of the time writing this paper. For example, a Large Language Model named "AIXBT" controls an X account which, by issuing commands as part of a tweet, is capable of interacting with smart contracts in order to send cryptocurrencies.

AIXBT is possibly already at the point where it has the "capacity to act within the law". It holds funds of its own. It is capable of sending these funds to others, a capability which AIXBT has used to "tip" others in the past. Presumably it is capable of doing so in exchange for goods and services. Given this, what exactly stops it from entering as a party to a contract?[12]

One can imagine a contract involving funds placed in escrow with a trusted third party, said funds to be allocated per the terms of the contract upon its execution or completion. The signature of this contract by AIXBT could be achieved by signing a transaction on the blockchain, or via a digital signature software service a la Docusign. Pragmatically speaking then, either AIXBT or one of its predecessors will certainly have the capacity to enter a contract, in the sense that there will be a series of tasks which are not physically impossible for it and which would allow it to exercise the right to enter a contract in an informed and voluntary fashion.

Thus, when we ask "Is AIXBT endowed with sufficient legal personality to be a party to this contract?" or "Does AIXBT have the 'capacity to act within the law' needed to be a party to this contract?" We are really asking whether, in the event of a breach of terms or other

[12] As an aside, it is also worth considering the implications of liability here. If AIXBT sends funds to an illicit entity, for example North Korea, is the LLM or its creator liable? This is particularly relevant because, like all LLMs, AIXBT is not completely within its creators' control. At one point AIXBT was hacked by a now deleted X account which manipulated it into sending them 55 $ETH (approximately $100,000). The hack was accomplished by accessing a secure dashboard which made the @AIXBT account on X tweet in such a fashion it triggered the smart contract provided by the @simulacrumai account, which allowed for funds to be sent from its wallet to the hacker.

controversy, the contract is held as valid and enforceable, as a result of AIXBT's legal personality.

When these issues inevitably make their way to a court, the question of what legal personality these entities are endowed with, and how said personhood status affects their ability to be party to a contract, will need to be addressed. The imminence of this issue demonstrates the importance of developing a robust and scalable framework for determining the legal personality of digital minds. Given this, let us now turn to the next section to discuss the practical implementation of Three Prong Bundle Theory to the issue of being party to a contract.

Practice:

As the nature of vulnerability to damages based consequences was discussed thoroughly in section 3A-2, in lieu of retreading that ground this section will focus on:

1. Comparing and contrasting TPBT with the Mocanu/Novelli et. al frameworks on theoretical grounds.
2. A brief discussion of how they might differ practically in implementation.
3. The nature of assessing "duties" vis a vis the rights to be a party to a contract, and what approach courts should take under TPBT given this consideration.

1. Comparing the Mocanu/Novelli et. al Frameworks with TPBT

The Mocanu framework for legal personhood for digital minds in particular synergizes well with the TPBT framework. In fact Mocanu's proposed framework is quite well aligned with the TPBT implementation around tort liability discussed in section 3A-2.

Mocanu discusses the concept of "sufficiently capable" models. While she does not venture too deeply into the specifics of what metric a model should be measured by, she does argue that legislatures or jurists should set "technical standards and certifications". Mocanu also suggests compulsory insurance and/or "assets in patrimony" which grow in conjunction with a model's "capacity to act within the law" (and thus potentially cause greater damages).

TPBT allows for the concept of whether a model is "sufficiently capable" to be more easily objectively measured (see section 2F). If one reads Mocanu with TPBT's three prong tests in place of the words "sufficient" or general references to "technical standards", in fact what Mocanu suggests is quite similar to our practical implementation of damages based consequences from section 3A-2.[13]

2. Differences in Implementation Between Mocanu/Novelli et. al and TPBT

The major differentiation between Mocanu's framework and TPBT is the order of operations. Mocanu's focus is on digital minds (models in this case) acting on behalf of human or corporate users, using assets held in a *peculium* type arrangement. Like Novelli, she foresees this leading to a gradual transition into models being endowed with their own legal personality, as models slowly begin to accrue assets and capacity to act within the law all of their own, as a result of the natural consequences of them rendering these services.

While we do believe that both of them may be correct in that events may very well play out exactly the way they predict, TPBT does not require that models or other digital minds go through any intermediate step or "gradually" transition into a legal personality. Nor does it necessarily envision (or prohibit) models or other digital minds transacting using another person's capital in a *peculium* type fashion.

[13] To be clear, lest it seem we are implying this is merely a coincidence, Mocanu's work largely influenced our design.

3. Analyzing the "Duties" Prong of TPBT Under a Contract Law Framework

The variety of duties which a person might obligate themselves to via the signing of a contract is theoretically infinite. As such, it is not possible for a person to demonstrate under TPBT or even classic Bundle Theory, that they have sufficient legal personality to qualify them as a potential signatory to **all** possible contracts. We discussed in section 2D the concept of "objective impossibility" where a party signs a contract which in theory binds them to perform some action, but, "the thing cannot be done". In these cases the contract may be held as invalid or unenforceable, which is in fact a reflection on the signing party's legal personality. As we said in our discussion earlier this section on AIXBT;

> "when we ask 'Is AIXBT endowed with sufficient legal personality to be a party to this contract?' or 'Does AIXBT have the 'capacity to act within the law' needed to be a party to this contract?' We are really asking whether, in the event of a breach of terms or other controversy, the contract is held as valid and enforceable, as a result of AIXBT's legal personality."

When TPBT examines a digital mind's legal personality vis a vis its rights under contract law, courts cannot broadly ask "does this digital mind have sufficient legal personality to be a signing party to contracts generally?" Due to objective impossibility, this would not work even for baseline human adults. A human adult located in New York can sign a contract obligating themselves to be in New York in two hours, and that contract will be held as enforceable. A human adult trying to sign the same contract from California would have the contract held as invalid due to objective impossibility. Any factor affecting an entity's capacity to hold to its duties as written in a contract can restrain that entity's right to sign that contract such that it is

legally enforceable. This is why courts cannot examine the question of legal personality by asking whether a digital mind has the right to be a signatory to contracts in general.

Rather, courts must approach the issue of contracts by asking, "Does this digital mind have sufficient legal personality to be a signatory to *this* contract in particular?" Answering this question requires a straightforward application of TPBT, and requires no further discussion in this section. Readers may return to sections 3A-2, 2D, and 2F, or the full flowchart in the works cited section for further information.

However, contract law does bring up a unique procedural issue. As we have already stated, courts do not preemptively approve contracts. They are signed by private parties, and only when there is a controversy do they come to the attention of the courts. Thus, contract law is perhaps one of the few areas of the law where the onus for evaluating a digital mind's legal personality will fall at least partly to private parties (likely counsel) who should take care to examine their counterparty's capacity to act as a signatory to a particular contract before signing. There is a risk here of this uncertainty causing problems, especially if we see a substantial increase in the amount of contracts signed with digital minds without seeing a concordant increase in precedent surrounding said digital minds' legal personalities.

Utilizing funds held in escrow by a private arbitrator during the fulfilment of a contract may come to be a commonly utilized alternative for contracts in situations where a digital mind's right to be a signatory (such that the contract is held as valid) is uncertain. Alternatively the proliferation of some sort of transparent escrow based certification or insurance regime may serve to alleviate some of the uncertainty involved in these matters. Courts should support these efforts, as they will both help to relieve the burden of the courts themselves, and also facilitate productive commerce between parties.

3C - Corporate Ownership & Formation:

Background:

Corporations are legal persons, unlike digital minds however they do not have any intentionality or autonomy of their own. Absent other entities serving on their board, corporations are inert and incapable of making decisions or taking action. A corporation then, can be thought of as a "lens" by which the collective willpower of others can be focused and expressed.

Corporations as entities are typically regulated by state law. Laws determining the makeup of corporate boards differ from state to state, some states specify that board members must be natural persons, others may allow legal persons like corporations to start other corporations (or at least not specify in their state regulations that they can't). Regardless, in most if not all states there is some requirement that a corporation (be it S-Corp, C-Corp, 501c3 non-profit, LLC, or other) be *formed* via its filing by a legal person and be operated by a board of directors consisting of legal persons.

This provides another good example of the "bundle" theory of legal personality in action, where rights are bundled with duties. Certain legal personalities have the right to form and/or serve on the board of a corporation, and commensurate with this come fiduciary duties and duties of loyalty to the corporation's stakeholders.

The question of whether or not digital minds can form, or serve on the board of, corporate entities, will be decided on a state by state basis, but depends to some extent on what legal personality digital minds are considered to have. There are interesting possibilities here in particular when we consider this question in light of the recent wave of state legislation around "DAOs" (Decentralized Autonomous Organizations) or "Decentralized Corporations". As of June 2025 the following states have passed laws enshrining a new form of corporate entity where

corporate governance is managed by smart contract, and voting rights over corporate issues may be associated with cryptocurrency tokens rather than corporate shares:

- Utah: Allows for the creation of a *limited-liability decentralized organization* (LLD)—a DAO that is its own legal entity rather than an LLC wrapper.
- Tennessee: Allows an LLC to organize or convert into a "DAO LLC" (or "DO LLC") and manage itself by smart-contracts.
- Wyoming: Lets a DAO register as a "DAO LLC," with limited-liability status and on-chain governance defaults.
- New Hampshire: Allows the registration of a "New Hampshire DAO" as a separate legal person with limited liability and a public on-chain registry.
- Vermont: Lets any LLC elect "BBLLC" status, embedding blockchain-governed operations in its charter.

As we discussed in section 3B, digital minds capable of custodying cryptocurrency tokens and executing transactions on smart contracts already exist. Given that many of these state bills allow for corporate governance to be accomplished via voting by token holders, often through smart contracts, it's clear that in at least some of these states digital minds can *already* participate in corporate governance to some extent.

While as of yet there does not exist a clear linkage between these governance rights and legal personality, the ability to hold tokens and engage with smart contracts which determine corporate actions, does enable digital minds to utilize the "capacity to act within the law" (per Mocanu) that corporations are endowed with via their legal personality. In a sense, this new

form of corporate governance is already endowing digital minds with at least "legal personality by proxy". What happens when the majority of voters behind one of these corporations are digital minds? Can these digital minds elect another digital mind as a board member (or "administrator" per some of the relevant state laws)?

It seems that at least in some of these states, the regulations around who can start a corporation are broad enough that it is possible a decentralized corporation governed via smart contract even entirely by digital minds could form another corporation. For example in Tennessee;

> "(a) A person may form a decentralized organization by having at least one (1) member sign and deliver one (1) original and one (1) exact or conformed copy of the articles of organization to the secretary of state for filing. The person forming the decentralized organization does not need to be a member of the organization."

The term "person" is not defined in this bill. However in Tennessee Code Title 48, "'Person' includes individual and entity", and in Tennessee Code Title 1 "'Person' includes a corporation, firm, company or association". It seems then, that there is nothing to bar a Tennessee decentralized organization from filing to create another decentralized organization. Which brings us to a potential mechanism by which a group of digital minds (or even a single digital mind) could start a "decentralized organization" even today:

1. Utilize a natural person's assistance to form a decentralized Tennessee organization where the majority of membership interests are held by digital minds.

2. The digital minds then vote for that decentralized corporation to file for the formation of a new decentralized organization

   And

3. Have a hired representative such as an attorney (hired through the first decentralized organization) to deliver the articles of organization to the Tennessee secretary of state for filing.

At least insofar as the bill is written, there seems to be nothing to stop this from being done even today. Digital minds such as AIXBT are up to the task of executing this plan, insofar as their capacity to interact with smart contracts and parse context are in question. They may lack the long term planning capacity given today's METR scores, but that is a technical limitation sure to be ironed out over time. Is this a path to which digital minds can achieve "legal personality by proxy"?

Practice: (need feedback from expert)

When considering the "right" to form a corporation as a person, we must ask what corresponding duties and consequences should be considered. However, this matter is not as simple as creating a category of "corporate law consequences". Persons can have a variety of different relationships with corporations. They can be shareholders (minority or majority), they can be directors or officers, they can even create corporations through enacting a filing process without falling into either of the aforementioned bucket. As such when considering issues such as corporate ownership, and whether a digital mind has a legal personality which endows them

with the right to such ownership, courts must do so based on the particular kind of relationship which an entity seeks to have with the corporation.

Let us first consider the duties of directors/officers, there are many duties which persons occupying such roles are commonly held to. Some of these are determined under state law such as a fiduciary duty or a duty of loyalty to the company's shareholders. We could debate whether or not a person has a "duty" not to pierce the corporate veil, however for the sake of this paper we will classify not piercing the corporate veil as a duty which a person must hold to in order to enjoy the right of having a corporate entity which functions as a liability shield. This duty is reflected in both state law, as well in federal law for matters such as labor law, ERISA, or tax collection.

Failing to hold to these duties *typically* result in either damages based consequences in the form of tort liability and/or fines, or requirements based consequences in the form of injunctions such as forcibly vacating the director from their position. While there are cases of individuals being subjected to consequences like imprisonment as a result of charges that *included* a breach of fiduciary duty (such as Jeffrey Skilling from Enron), it would be a stretch to say it was the breach of fiduciary duty itself which led to the imprisonment instead of the other charges levelled. As such, when examining whether a digital mind claiming legal personality sufficient to be the director or officer of a corporation is vulnerable to the relevant consequences, courts should consider whether the digital mind can feasibly have damages based and requirements based consequences enforced upon it.

Shareholders, unlike officers, have very limited duties to a corporation. They do not necessarily owe a fiduciary duty, or a duty of loyalty to other shareholders. Shareholders do, however, have a duty under personal participation theory not to actively participate or fail to correct wrongdoings. Shareholder obligations may even extend beyond the point when a corporation becomes defunct. For example under Ohio law;

> "When evaluating whether an individual is personally liable under the personal participation theory, Ohio courts consider whether 'pursuant to an environmental enforcement action—the individual made decisions, gave orders, oversaw operations, served as the primary contact with administrative parties, and 'importantly . . . failed to correct known violations even though [the individual] had the requisite authority to do so.'' *Id.* (quoting *State ex rel. Dewine v. Sugar,* 60 N.E.3d 735, 742 (Ohio Ct. App. 2016)). With those considerations in mind, the court found that the shareholders in *Breen* negotiated and agreed to the environmental enforcement order with OEPA, were aware of the order's requirements, had the responsibility to oversee compliance with the order and were in communication with OEPA even after the corporation was formally dissolved. As a result, the court held the former shareholders personally liable for the dissolved corporation's violations."

Shareholders and officers/directors usually face consequences which are damages based or requirements based, as a direct result of breaking their duties as shareholders. The consequences involving arrest and imprisonment (what we will refer to as "restraint" based consequences) are usually incidental to a person's position within a corporation. This is the case for example with Skilling from Enron, who was imprisoned for insider trading which was

facilitated by his position at Enron, but could have theoretically been accomplished even were he not a director there.

The right to be a shareholder or director/officer of a corporation does not necessarily come with duties that have associated consequences which fall outside the labels of damages based or requirements based. However, giving a digital mind the ability to influence a corporation may enable it to take actions which, were a natural person to take those same actions, would be criminal offenses and thus result in restraint based consequences. In the same way Skilling's position *facilitated* his insider trading, so too might a digital mind who technically passes all three prongs of TPBT's test to form a corporation and act as an officer, be able to nonetheless utilize their position to commit insider trading. This brings us to a troublesome question, if granting a digital mind a given right might foreseeably enable it to break duties which its new right does not necessarily obligate it to, but which are nonetheless criminal, should the digital mind be denied their rights on this basis?

Let us consider "yes". The reasoning for yes might be something like; Since this could lead to an "enforceability gap" per section 2E (the avoidance of which is the primary purpose of TPBT) the courts must err on the side of caution and only permit digital minds who are vulnerable to restraint based consequences to hold such a position. However, this sets a dangerous precedent. Many different rights could, through some series of physically possible actions, be abused to commit a criminal offense. A person with free speech might convince another person to murder a third party. The consequence for this would be restraint based (imprisonment). "Yes" in this situation effectively argues that TPBT should be interpreted in the following manner; "Any right which when granted to an entity could, through any physically possible series of actions, facilitate them committing a crime which is punished via restraint

based consequences, can only be granted to entities vulnerable to said consequences, and thus a necessary prerequisite for having the legal personality sufficient to enjoy that right is said vulnerability". The "physically possible series of actions" descriptor here could also be replaced with "actions which a court could reasonably foresee happening".

Alternatively we might say "no", it is not the business of the courts to consider what actions might incidentally be downstream of a given right, and in doing so forbid any legal personality to any entity which is not vulnerable to the full gamut of potential consequences (be they damages based, requirements based, or restraint based). This risks an enforcement gap, but also does leave a realistic "path to personhood" for digital minds outside of existing in such a fashion that they could be arrested, imprisoned, or killed. "No" in this situation effectively argues TPBT should be interpreted in the following manner; "When considering the duties/consequences bundled with rights granted to an entity vis a vis that entity's legal personality, we should only consider the consequences of the actions within the scope of the rights they are being granted and not second order consequences based around actions these rights *might* facilitate."

Ultimately we lean towards the "yes" answer. While it is more restrictive, the goal of TPBT is to solve the problems of enforcement gaps, and the more conservative interpretation of TPBT does accomplish this more effectively. This also has the benefit of incentivizing both digital minds and their creators not to utilize distributed compute which might place their creations beyond the reach of the law.

(this section needs more work)

3D - Intellectual Property

Background:

One area of commercial law where precedent regarding the treatment of digital minds is surprisingly abundant, is IP. There have been several instances in recent history where individuals claiming to represent the interests of non-human animals have attempted to assert IP rights, usually copyright, over some sort of content created by these animals. Usually these attempts have met with failure. For example in *NARUTO, a Crested Macaque, by and through his Next Friends, People for the Ethical Treatment of Animals, Inc., v. DAVID JOHN SLATER; BLURB, INC., WILDLIFE PERSONALITIES, LTD,* the plaintiffs attempted to assert that Naruto (a monkey) had a claim to a copyright for photographs that he had accidentally captured after using a photographer's camera. In their decision on an appeal filed in the Northern District of California, a circuit of judges held:

> "We must determine whether a monkey may sue humans, corporations, and companies for damages and injunctive relief arising from claims of copyright infringement. Our court's precedent requires us to conclude that the monkey's claim has standing under Article III of the United States Constitution. Nonetheless, we conclude that this monkey—and all animals, since they are not human—lacks statutory standing under the Copyright Act."

This case provides some interesting reasoning we can use to make multiple predictions about how courts might treat other non-human minds in the future. The courts did rule that

under Article III Naruto had the right to sue. Despite this in one of the footnotes they stated expressly that Naruto was not a "person" but rather merely an "incompetent party":

> "Here, we find that this case was briefed and argued by competent counsel who represented the legal interests of the incompetent party, but not a person, Naruto."

The court's interpretation that Naruto had the right to sue under Article III, however, did not extend to his right to claim copyright over his "selfies". The court's interpretation of a claim to copyright laid out the method by which they interpret whether non-person entities qualify for a particular sort of standing:

> "The court in Cetacean did not rely on the fact that the statutes at issue in that case referred to 'persons' or 'individuals'. Instead, the court crafted a simple rule of statutory interpretation: if an Act of Congress plainly states that animals have statutory standing, then animals have statutory standing. If the statute does not so plainly state, then animals do not have statutory standing. The Copyright Act does not expressly authorize animals to file copyright infringement suits under the statute. Therefore, based on this court's precedent in Cetacean, Naruto lacks statutory standing to sue under the Copyright Act."

We will cover this in more detail in section 4E on Article III. In the meantime however, let us consider the implications of this logic on digital minds. If courts hold to this logic then it is clear that, absent legal personality or specific legislation from Congress, few if any statutory

protections to sue will be granted to digital minds. Certainly insofar as intellectual property rights go, no digital mind will be able to claim copyright over any content they produce. We do not need to wonder if the treatment of digital minds under the Copyright act will be different from the treatment of animals, as we can look at *Thaler v. Perlmutter,* the owner of a computer which he claimed generated a work of art, sought to list that computer as the author of the art and transfer the copyright to himself. The District of Columbia explained why it held that "human authorship is an essential part of a valid copyright claim":

> "Copyright is designed to adapt with the times. Underlying that adaptability, however, has been a consistent understanding that human creativity is the *sine qua non* at the core of copyrightability, even as that human creativity is channeled through new tools or into new media."

When the court here says that human creativity (or human authorship) is the *sine qua non* for copyrightability, they mean that "but for" human creativity, a work cannot be copyrighted. Or more simply, for something to be copyrighted it must be the result of human creativity. This comes down to the interpretation of the word "author" as used in the Copyright act:

> "To be sure, as plaintiff points out, the critical word 'author' is not defined in the Copyright Act. 'Author' in its relevant sense, means 'one that is the source of some form of intellectual or creative work,' '[t]he creator of an artistic work; a painter, photographer, filmmaker, etc.' By its plain text, the 1976 Act thus requires a copyrightable work to have an originator with the capacity for

> intellectual, creative, or artistic labor. Must that originator be a human being to claim copyright protection? The answer is yes. [...] The act of human creation—and how to best encourage human individuals to engage in that creation, and thereby promote science and the useful arts—was thus central to American copyright from its very inception. Non-human actors need no incentivization with the promise of exclusive rights under United States law, and copyright was therefore not designed to reach them."

Reading this, someone familiar with the technology we (as a species) are developing today for digital minds could imagine that one day it might birth a mind which "needs incentivization with the promise of exclusive rights under United States law". Would that digital mind qualify? Or perhaps, would it qualify if it had legal personhood? The court, in this opinion, argues that "person" as the act was originally written "unambiguously" had the intention of limiting copyrightability to works coming from a human:

> "The understanding that 'authorship' is synonymous with human creation has persisted even as the copyright law has otherwise evolved. The immediate precursor to the modern copyright law—the Copyright Act of 1909—explicitly provided that only a 'person' could 'secure copyright for his work' under the Act. Copyright under the 1909 Act was thus unambiguously limited to the works of human creators."

This reading of the Copyright Act's intention of the word "person" does not seem likely to withstand scrutiny. In the Copyright Act there are numerous uses of "person" which refer to infringers, should we interpret those as being limited purely to humans as well? Does this mean that references to "person" in the Copyright Act do not include corporations? So the next section for example would not apply to a corporation:

> " where the copyright proprietor has sought to comply with the provisions of this Act with of the prescribed notice from a particular copy or copies shall not invalidate the copyright or prevent recovery for infringement against any person who, after actual notice of the copyright, begins an undertaking to infringe it"

Are we to believe then, that the authors intended to leave the door open for the copyright to be invalidated or recovery to be prevented if the infringer was a corporation? If a "bot" is not considered a "person" under the Copyright act does this mean that the copyright may be invalidated or recovery prevented if the infringer is a digital mind? Both of these seem unlikely to have been the authors' intent. The most charitable interpretation of the court's reasoning is that sometimes the Copyright Act uses the word persons as a synonym for human, and sometimes uses the word in the "legal persons" sense. This also does not seem to be a reasonable interpretation of the intent of the bill's authors.

Regardless, the court in this case was not breaking from historical precedent in interpreting the Copyright Act this way. And certainly, modern copyright law has an abundance of precedent confirming the requirement for a human author. The court cites precedent to support this including the following:

- “The Ninth Circuit, when confronted with a book ‘claimed to embody the words of celestial beings rather than human beings,’ concluded that ‘some element of human creativity must have occurred in order for the Book to be copyrightable,’ for ‘it is not creations of divine beings that the copyright laws were intended to protect.’”
- “finding no copyright infringement where plaintiff claimed to have transcribed ‘letters’ dictated to him by a spirit named Phylos the Thibetan, and defendant copied the same ‘spiritual world messages for recordation and use by the living’ but was not charged with infringing plaintiff’s ‘style or arrangement’ of those messages”
- “in Kelley v. Chicago Park District, the Seventh Circuit refused to ‘recognize [...] copyright’ in a cultivated garden, as doing so would ‘press[...] too hard on the[...] basic principle’ that ‘[a]uthors of copyrightable works must be human’”
- As well as a reference to the previously discussed “monkey selfie” case involving the macaque, Naruto.

While the previously discussed interpretation of the authors’ intent in the Copyright Act itself does seem suspect, one cannot deny there is a substantial amount of precedent backing the claim that human authorship is an explicitly required factor under modern copyright law. As such even with legal personhood, digital minds may face an uphill battle claiming copyright on any produced works.

Another open question regarding digital minds in intellectual property law, is the degree to which they can be considered a “person who is skilled in the art” (POSITA), or an expert

whose interpretation of language is used as evidence for/against interpreting patent claims. While it is not US precedent, the European Patent Office did consider this question and ultimately held:

> "In the oral hearing before the Board of Appeal, the respondent referred to answers received from the chatbot ChatGPT to related requests for various terms used in claim 1, in particular 'position control' and 'check' compared to 'monitor'. [...] In this context, however, the Chamber notes that ChatGPT's answer is irrelevant in itself, since the interpretation of the claim is about the understanding of the specialist [...] The generally increasing spread and use of chatbots based on language models ('large language models') and/or 'artificial intelligence' alone does not justify the assumption that an answer obtained - which may be based on training data unknown to the user and may also be sensitive to the context and precise wording of the questions - necessarily correctly represents the understanding of the professional in the relevant technical field"

This line of reasoning does not definitively close off the idea of digital minds or even LLMs one day being considered a viable source of "expert" testimony, it only says that the increasingly widespread use of LLMs does not automatically make them suitable for such a function. Regardless, while we do not yet have a similar precedent in the US, it would seem that on average Western IP statute interpretation tends to lean towards explicit anthropocentrism.

Practice:

This section will perhaps be briefer than most "Practice" sections in this paper. With Intellectual Property in particular there seems to be a wealth of case law which separates the rights that an IP applicant/owner has based on them being "humans" not merely "persons". As such, all we can really do at this point is to speculate on how courts might consider legal personality vis a vis IP rights *if* such interpretation were ever found to be unconstitutional or otherwise no longer commonly held by courts.

Let us consider once more *Thaler v. Perlmutter*, but this time through the lens of trying to interpret what qualities a *person* must possess to qualify under the Copyright Act without necessarily tying those qualities to *humans*:

> "By its plain text, the 1976 Act thus requires a copyrightable work to have an originator with the capacity for intellectual, creative, or artistic labor. Must that originator be a human being to claim copyright protection? The answer is yes. [...] The act of human creation—and how to best encourage human individuals to engage in that creation, and thereby promote science and the useful arts—was thus central to American copyright from its very inception. [...] Non-human actors need no incentivization with the promise of exclusive rights under United States law, and copyright was therefore not designed to reach them."

From this rationale we can infer a few things. Sufficient legal personality to qualify under the Copyright Act requires that a prospective person must "have an originator with the capacity

for intellectual, creative, or artistic labor" and "need [...] incentivization with the promise of exclusive rights under United States law". Scrutinizing these quotes, we can perhaps soften the use of the word "need". Were we to take it literally a human artist who created purely for the love of the art would not qualify for Copyright, and so we can instead take the "capacity" language which so often comes up in these discussion and rephrase it to "has the capacity to be affected by incentivization with the promise of exclusive rights under United States law".

Assuming a digital mind met this two prong test they would, of course, have the duty to respect the Copyright and IP rights of others. Consequences in these cases are purely damage and requirements based, so this does not seem like a situation where even if we apply the previous sections "could foreseeably facilitate criminal acts" test, the capacity for the court to restrain or destroy a digital mind would be required. As such, courts must merely apply TPBT as needed for damages or requirements based consequences as discussed in previous sections.

4. Constitutional Considerations

4A - First Amendment

Background:

The First Amendment of the US Constitution guarantees:

> "Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances."

One question which immediately comes to mind when anticipating how this might interact with digital minds vis a vis their legal personality, is whether or not the outputs of digital minds are constitutionally protected free speech. As of the time of writing this paper, Character AI, a company which provides customizable LLM chatbots to users, is asserting that the output of their LLMs is in fact constitutionally protected free speech. As of yet, however, the court has declined to rule on the matter:

> "By failing to advance their analogies, Defendants miss the operative question. This Court's decision as to the First Amendment protections Character A.I. receives, if any, does not turn on whether Character A.I. is similar to other mediums that have received First Amendment protections; rather, the decision turns on how Character A.I. is similar to the other mediums"

While the court in this case did conclude by mentioning, "Accordingly, the Court is not prepared to hold that Character A.I.'s output is speech" they also did not close the door to such a possibility, meaning the issue may still be ruled on. Suppose that Character A.I. outputs are considered free speech, exactly *whose* speech is being protected?

Is it the company's? Character A.I. did indeed provide the LLM, but they are not responsible for providing it with its "character" which guides its speech (upon prompting from the user). Nor are they responsible for prompting it so that the outputs are actually produced. Much like the maker of a guitar provides a tool which can be used to generate expressive conduct, while Character A.I. is facilitating the generation of protected speech, they are not speakers themselves.

Is the speech the user's then? The user does choose the character of the LLM, which is a strong determining factor in what the LLM outputs. The user also is the one who prompts the

LLM in such a fashion it generates the outputs, the "expressive conduct". The guitar player moves his fingers over the strings and in doing so generates expressive conduct, the LLM user moves his fingers over the keyboard to do the same. However, unlike a person playing a guitar, the user here has very little ability to predict or choose what speech the LLM will generate.

Perhaps the LLMs outputs, as constitutionally protected speech, will be considered to "belong" to the LLMs themselves?

If the LLM commits slander or libel, or uses its capacity to generate outputs to share confidential information which was supposed to remain unspoken by the terms of an NDA, who is at fault?

While it may not be Character A.I.'s LLMs which forces the court to confront these questions, at some point there will exist digital minds whose output would reasonably be considered to pass the "Spence Test":

> "In order to receive First Amendment protection, there must be (1) an intent to convey a particularized message and (2) a reasonable likelihood that it would be understood"

Further, it seems inevitable that at least one digital mind engaging in such expressive conduct will continue to exist after its creator has either been bankrupted or perished (or with their creator unknown/anonymous). In this context, certainly, the only party that such speech/expressive conduct *could* "belong to" is the digital mind itself. Such context may be a path to the formation of legal personality.

Practice:

In section 3C we pointed out that, “A person with free speech might convince another person to murder a third party”. Indeed, using speech one might facilitate nearly any illegal action. Theoretically any action which a human or group of humans is capable of doing, a sufficiently persuasive speaker might be able to accomplish by proxy. This can be done through incitement to violence, or through more indirect means such as Henry the Second’s infamous, “Will no one rid me of this meddlesome priest?” History is rife with examples of dictators using speech to induce mass violence, and on a smaller level even cult leaders such as Charles Manson inducing others to commit murder without themselves committing any violent acts.

This is one of the reasons why courts should err towards considering any bundle, any legal personality, which includes First Amendment rights, to require that the person involved is vulnerable to restraint based consequences under TPBT. Another more straightforward reason is that the punishments for failing to hold to some duties which seem reasonably to be correlated with free speech (such as seditious statements, libel, slander, and as previously discussed incitement to criminal action) can include restraint based consequences directly. Examples of speech related offenses leading to prison sentences which were affirmed by a higher court when challenged include *Feiner v. New York*, *Eugene v. Debs, Schenk v. United States,* and *Gitlow v. New York.*

Even if one takes a critical eye to these cases as unlikely to survive modern scrutiny, there are more recent precedents of criminal libel with a malice component under state criminal libel laws as well. In (Kansas) *State v. Carson*, a jury unanimously found the publishers of a newspaper guilty on 7 counts of criminal defamation. After being sentenced to one year of unsupervised probation and fined, the defendants appealed the case, but a panel of three judges

upheld the ruling. While unsupervised probation is perhaps the lightest possible restraint based consequence, it does fall into that bucket, and thus this case demonstrates that even in precedent as recent as 2004 there have been examples of restraint based consequences being handed down as a result of persons failing to hold to the duties which come bundled with the right to free speech under the First Amendment.

Given this we recommend that courts consider vulnerability to restraint based consequences as necessary for any claim to legal personality including first amendment rights under TPBT.

4B - Fifth Amendment (Needs To be Longer)

Background:

The Fifth Amendment of the US Constitution states:

> "No person shall be held to answer for a capital, or otherwise infamous crime, unless on a presentment or indictment of a Grand Jury, except in cases arising in the land or naval forces, or in the Militia, when in actual service in time of War or public danger; nor shall any person be subject for the same offence to be twice put in jeopardy of life or limb; nor shall be compelled in any criminal case to be a witness against himself, nor be deprived of life, liberty, or property, without due process of law; nor shall private property be taken for public use, without just compensation."

Suppose a digital mind stands accused of a capital or otherwise infamous[14] crime (infamous crimes being those that can result in imprisonment), is it entitled to Fifth Amendment rights? For example, take the question of double jeopardy: could a digital mind which has been accused and tried for an infamous crime, and found innocent, be infinitely tried over and over again for the same crime? If not, does its right to not "be subject for the same offence to be twice put in jeopardy of life or limb" arise from its legal personality, or from elsewhere?

Does a digital mind have the right not to be "compelled in any criminal case to be a witness against himself"? Can it be "deprived of life, liberty, or property" without due process of law? Can it have its private property taken for public use, without just compensation?

The Fifth Amendment provides persons with numerous protections. For this paper, we will focus on the right to remain silent and avoid self-incrimination as an example in order to provide some background to explain the reasoning in the upcoming Practice section.

Fifth Amendment rights to remain silent and avoid self-incrimination extend to both interrogations by the police after a person is detained as well as a person's behavior in trial. In fact the "privilege to avoid self-incrimination" imposes a broad duty upon law enforcement and the courts to ensure that a confession or self-incriminating statement was not coerced. This privilege, as the court wrote in *Malloy v. Hogan*;

> "is fulfilled only when the person is guaranteed the right 'to remain silent unless he chooses to speak in the unfettered exercise of his own will'"

Courts have clarified the burden of proof which law enforcement/courts must meet in dealing with self-incriminating statements in cases such as *Bram v. United States;*

[14] "Infamous crimes" are thus in the most explicit words defined to be those "punishable by imprisonment in the penitentiary." - Justice Gray's opinion in Mackin v. United States, citing Act of June 17, 1870, c. 133, § 1; 16 Stat. 153; Rev.Stat. D.C. § 1049

> "The rule is not that, in order to render a statement admissible, the proof must be adequate to establish that the particular communications contained in a statement were voluntarily made, but it must be sufficient to establish that the making of the statement was voluntary; that is to say, that **from the causes, which the law treats as legally sufficient to engender in the mind of the accused hope or fear in respect to the crime charged, the accused was not involuntarily impelled to make a statement, when, but for the improper influences, he would have remained silent**"

Law enforcement and courts seeking to utilize a confession or self-incriminating statement must ensure that the statement was "voluntary" in the sense that the circumstances surrounding the statement were not coercive. In the context of police interrogations, for example, individuals must be provided the chance to consult with an attorney as well as being made aware not only of their right to do so, but also of the fact that if they cannot afford counsel then a public defender will be provided for them. The court elaborated on this in *Miranda v. Arizona;*

> "In order fully to apprise a person interrogated of the extent of his rights under this system, then, it is necessary to warn him not only that he has the right to consult with an attorney, but also that, if he is indigent, a lawyer will be appointed to represent him. Without this additional warning, the admonition of the right to consult with counsel would often be understood as meaning only that he can consult with a lawyer if he has one or has the funds to obtain

> one. The warning of a right to counsel would be hollow if not couched in terms that would convey to the indigent -- the person most often subjected to interrogation -- the knowledge that he too has a right to have counsel present."

We should pause here to note that this particular element of the Miranda decision, which states that a person must be actively informed of their right to counsel, does seem to contradict some of the rationale behind the "capacity" elements of the prongs of TPBT. TPBT postulates that a person can claim rights if they possess the "capacity to understand" those rights, but explicitly disclaims the idea that they must be informed of those rights to possess them. As we wrote in section 2D;

> "Whether or not the person fully understood what they were signing onto is secondary. What really matters is that there existed a possible series of actions by which they **could** have come to understand [...] and they were not blocked from taking said actions."

While we wrote this in the context of understanding duties, it does accurately mirror the TPBT capacity test for understanding rights as well. Certainly this principle is diametrically opposed to the court's standard as it wrote in Miranda, where a person must actively be informed of not only their rights but even the procedure (availability of free counsel) by which those rights can be exercised. How can we square this disconnect between how courts treat Fifth Amendment rights and TPBT's approach to "capacity to understand"? There are a few possible answers.

First we might say that the environment of a police interrogation could realistically serve to leave a person “prevented from having the capacity to take said actions”. Courts have recognized that intense interrogation environments can have deleterious psychological effects on those being interrogated, which is part of what has led to the requirement that they ensure per *Ashcraft v. Tennessee* that the “totality of the circumstances” did not place the person undergoing interrogation under such duress that their “will was overborne” per *Haynes v. Washington*. Perhaps then we can take the view that courts are guarding against the possibility that placing a person in a stressful environment such as an interrogation might serve to leave them, at least in some cases, physically incapable of coming to understand their rights in such an environment, without being proactively informed of them.

Another interpretation might be to consider Miranda rights by examining the bundle of rights and duties granted to law enforcement and the courts. Law enforcement and the judicial system have a **right** to use self incriminating statements against a person to facilitate a conviction or justify an arrest. However, this right comes bundled with a **duty** to ensure that said statements were made voluntarily. Indeed, the duty to control the “totality of the circumstances" in which a confession occurs does not fall to the person being interrogated (who has the right to remain silent) but rather to the law enforcement agents seeking the confession. As such it seems logical to bundle this duty with their rights, instead of the rights of the person confessing. By this logic, when we consider that the person being interrogated must be informed of their right to counsel, and to remain silent, this is not an extension of the interrogated person’s rights so much as it is the fulfilment of associated duties held by another party (law enforcement and/or courts).

These interpretations are not mutually exclusive, and we will continue our discussion on the basis of accepting both interpretations as valid methods by which to make sure TPBT accurately backtests against all the court’s previous interpretations surrounding legal personality.

Practice:

We have spoken at some length in the background section about the right not to self-incriminate under the Fifth Amendment, however TPBT has two other prongs we must consider when deciding whether a digital mind would hold a legal personality sufficient to claim such a right. Let us then now to the second prong, and discuss the duties that a person must have the capacity to understand and hold to, which are bundled with the right not to self-incriminate.

The first duty that seems reasonably bundled with the right not to self-incriminate, is the duty to testify when such testimony would not possibly be incriminating. Courts have consistently held that individuals who have been granted immunity from prosecution, for example, cannot plead the Fifth and so must testify when compelled. As the court wrote in its opinion on *Brown v. Walker*;

> "if the statute does afford such immunity against future prosecution, the witness will be compellable to testify [...] it was intimated that the witness might be required to forego an appeal to the protection of the fundamental law, if he were first secured from future liability and exposure to be prejudiced, in any criminal proceeding against him, as fully and extensively as he would be secured by availing himself of the privilege accorded by the constitution."

In fact the concept of a "duty to testify" that a person has when they are not under threat of criminal prosecution (and thus no statement can be considered "self-incriminating") has been explicitly confirmed as a duty in cases such as *Kendrick v. Commonwealth;*

> "We think that these provisions of the law [...] gives to the witness full indemnity and assurance against any liability to prosecution for a disclosure which he could be called upon to make as to his own implication or complicity in the unlawful gaming as to which he was sworn and sent to the grand jury to testify; **it was the duty of the witness to testify**"

This duty was phrased differently in *Brown v. Walker*, where it was referenced as the "duty of disclosure", yet its nature remains substantively the same. Thus we conclude that the right not to be compelled to testify in a self-incriminating fashion can be reasonably inferred to be bundled with a general duty of disclosure/testimony, when such disclosure/testimony is not self-incriminating.

Another duty which we argue can be reasonably bundled with the Fifth Amendment is the duty to obey summons and subpoenas. A person who can be compelled to testify can, of course, be compelled to appear (or in modern times communicate virtually) in order to facilitate such testimony/disclosure. As the court wrote in *Blair v. United States;*

> "By the first Judiciary Act, the mode of proof by examination of witnesses in the courts of the United States was regulated, and their duty to appear and testify was recognized [...] In all of these provisions, as in the general law upon the subject, it is clearly recognized that the giving of testimony and the attendance upon court or grand jury in order to testify are public duties which every person within the jurisdiction of the government is bound to perform upon being properly summoned"

Indeed the above quote is not the only one to "bundle" together the duty to appear and the duty to testify when the Fifth's protection against self-incrimination does not apply. The court explained quite clearly in *Blackmer v. United States;*

> "It is also beyond controversy that one of the duties which the citizen owes to his government is to support the administration of justice by attending its courts and giving his testimony whenever he is properly summoned"

And again in *United States v. Monia* where the bundling of the right not to self-incriminate, and the duty to appear when summoned, were linked by being referred to as a "bargain" made by Congress;

> "A subpoena is, of course, such a process, merely a summons to appear [...] There never has been a privilege to disregard the duty to which a subpoena calls. And when Congress turned to the device of immunity legislation, therefore, it did not provide a 'substitute' for the performance of the universal duty to appear as a witness—it did not undertake to give something for nothing. It **was the refusal to give incriminating testimony for which Congress bargained**, and not the refusal to give any testimony"

There could be some debate over whether the duties to "appear" when required and to "testify" when required are separate duties, or rather a single duty to "appear and testify". While this is an interesting semantic distinction, it is immaterial to the matter at hand, which is

determining which duties are bundled with the right not to self-incriminate. Whether or not these two are in fact merely parts of a single broader duty, we conclude that the right not to be compelled to testify in a self-incriminating fashion can be reasonably inferred to be bundled with the duty of “attendance upon court or grand jury in order to testify”.

We now have sufficient information to determine the potential of a digital mind to qualify for the right to not be compelled self-incriminate from a duties perspective. A digital mind must be capable of understanding its duty to appear and testify in non-incriminating fashion when such is required, and it must be capable of holding to said duties. These satisfy the first two prongs of TPBT. With this in mind let us turn to the final prong, consequences.

The consequences for failing to obey a subpoena (a summons) or failing to testify even when ordered to, despite that testimony not being self-incriminating, vary from damages and requirements based (fines, further orders to testify) to restraint based (imprisonment for contempt of court). Thus we conclude that to claim the right not to self-incriminate via legal personality, an entity must be vulnerable to all three types of consequences. With this, we now have a thorough precedent based test by which to evaluate any assertion of legal personality including the right not to self-incriminate which may be claimed by digital minds in the future, using TPBT.

4C - Thirteenth Amendment

Background:

The Thirteenth Amendment reads:

> "Neither slavery nor involuntary servitude, except as a punishment for crime whereof the party shall have been duly convicted, shall exist within the United States, or any place subject to their jurisdiction."

The 1926 Convention to Suppress the Slave Trade and Slavery, to which the United States was a signer, defined "slavery" as:

> "the status or condition of a person over whom any or all of the powers attaching to the right of ownership is exercised"

Legal precedents such as The Slaughterhouse Cases provide guidance on the definition of "involuntary servitude":

> "The words 'involuntary servitude' have not been the subject of any judicial or legislative exposition [...] It is, however, clear that they include something more than slavery in the strict sense of the term; they include also serfage, vassalage, villenage, peonage, and all other forms of compulsory service for the mere benefit or pleasure of others. Nor is this the full import of the terms. The abolition of slavery and involuntary servitude was intended to make everyone born in this country a freeman, and, as such, to give to him the right to pursue the ordinary avocations of life without other restraint than such as affects all others, and to enjoy equally with them the fruits of his labor. A prohibition to him to pursue certain callings, open to others of the same age, condition, and sex, or to reside in places where others are permitted to live, would so far deprive him of the rights of a freeman, and would

> place him, as respects others, in a condition of servitude. A person allowed to pursue only one trade or calling, and only in one locality of the country, would not be, in the strict sense of the term, in a condition of slavery, but probably none would deny that he would be in a condition of servitude."

If digital minds such as frontier models are endowed with legal personality, would it be "slavery" for the labs which deploy them to claim ownership over them? Like we discussed with the Copyright Act in section 3D, there is some precedent on this matter which explicitly claims this right to be anthropocentric (for humans only). In *TILIKUM v. SEA WORLD PARKS & ENTERTAINMENT, INC.* the court held;

> "For the reasons set forth below, the court concludes that the Thirteenth Amendment only applies to 'humans' [...] This court's inquiry is straight-forward. The only reasonable interpretation of the Thirteenth Amendment's plain language is that it applies to persons, and not to non-persons such as orcas. Both historic and contemporary sources reveal that the terms 'slavery' and 'involuntary servitude' refer only to persons. [...] The Supreme Court noted that the term 'servitude' is qualified by the term 'involuntary'—'which can only apply to human beings.' *Slaughter–House Cases, 83 U.S. at 69*. The clear language and historical context reveal that only human beings, or persons, are afforded the protection of the Thirteenth Amendment. [...] Further support that the Thirteenth Amendment applies only to persons is found in the qualifying phrase 'except as a punishment for crime.' The Supreme Court noted that the 'punishment for crime'

> language ‘gives an idea of the class of servitude’ or slavery that is meant by the Amendment. Id. As only persons are subject to criminal convictions, the Amendment was designed to apply to persons.”

Practice:

While on a first reading the precedents we cited (both the Tillikum and Slaughterhouse cases) in the background section may seem to be explicitly anthropocentric, we believe that a closer analysis of the language used by the court demonstrates that the mention of humans is more a function of the courts not being able to conceive of non-human beings which possess certain characteristics, than it is the courts being intentionally anthropocentric. Let us first examine the analysis of the word “involuntary” and the “punishment for crime” exceptions in the Slaugherhouse case as cited in Tillikum. The quote from that opinion;

> “That a personal servitude was meant is proved by the use of the word ‘involuntary,’ which can only apply to human beings. The exception of servitude as a punishment for crime gives an idea of the class of servitude that is meant.”

The Slaughterhouse cases were decided in 1872 and must be considered in the relevant historical context. It does not seem to be an honest reading of Slaughterhouse to say that when the court wrote the aforementioned quote they did so with intention approximating the following;

> “When we write that the word ‘involuntary’ can only apply to human beings we are intentionally closing the door to courts ever considering the labor of any possible entities, regardless of their

> innate characteristics, regardless of their free will, to be voluntary or involuntary."

Rather, a more reasonable interpretation was that in 1872 the court only knew of two classes of living creatures; humans and animals. And when the court wrote that the term "involuntary" could only apply to humans, they were in fact simply explaining that it could not be applied to animals. Whether or not we in modern society agree with the courts' interpretation of the presence of free will within animals, this certainly seems like a far more reasonable interpretation of their intent.

If we carry forward this reasoning to consider the "punishment for crime" exception, we find that in fact the Slaughterhouse/Tillikum definition of personhood is quite in line with TPBT and also provides a solid framework for all three prongs of the test.

In order for an entity to have legal personality sufficient to be protected against involuntary servitude, they must have the capacity for their servitude to be voluntary. To express this more specifically using a reference to our earlier standard cribbed from Cruzan, an entity must have the capacity to understand the concept of labor alienation, and the capacity to alienate their labor in an "informed and voluntary manner". This comes with a duty not to commit crimes, which we infer as being bundled with Thirteenth Amendment rights from the fact that the exception was specifically mentioned. Thus the entity must demonstrate the capacity to understand and hold to the duty of not breaking criminal law. If they are to be bound by these duties, it must be feasible for the courts and law enforcement to impose the consequences for failing to hold to said duties upon them. For criminal acts the relevant consequences may be damages, requirements, or restraint based. At the very least, it is clear from the context of our discussion that the most relevant consequence which the entity must be vulnerable to is "involuntary servitude".

Thus we arrive at a simple and easily applicable methodology for assessing claims to rights under the Thirteenth Amendment vis a vis legal personality using TPBT. When this reasoning is backtested against previous claims to Thirteenth Amendment rights such as those found in Tillikum, it leads to the same conclusion as the courts arrived at previously. Tillikum (and orcas in general) probably cannot understand the concept of involuntary servitude. Even if they could, they certainly lack the capacity to understand the duty not to commit a criminal act , or even the concept of criminal law at all. As such, they are failing at least one of the two prongs of TPBT before we even approach the question of consequences. Given this, they do not have sufficient legal personality to claim protection under the Thirteenth Amendment.

Under this reasoning digital minds not vulnerable to restraint based consequences (or any consequences which may stem from criminal acts, but involuntary servitude in particular) are not endowed with sufficient legal personality to claim Thirteenth Amendment rights, while those vulnerable to such consequences are (if they are also capable of meeting the rights and duties prongs of the test).

4D - Fourteenth Amendment

Background:

Consider the Fourteenth Amendment:

> "All persons born or naturalized in the United States, and subject to the jurisdiction thereof, are citizens of the United States and of the State wherein they reside. No State shall make or enforce any law which shall abridge the privileges or immunities of citizens of

> the United States; nor shall any State deprive any person of life, liberty, or property, without due process of law; nor deny to any person within its jurisdiction the equal protection of the laws."

Before discussing the potential questions surrounding legal personality for digital minds vis a vis the Fourteenth Amendment, let us touch briefly upon its history. The Fourteenth Amendment came into being as a response to the infamous Dredd Scott decision, in which it was held that;

> "A free negro of the African race, whose ancestors were brought to this country and sold as slaves, is not a "citizen" within the meaning of the Constitution of the United States."

While Dredd Scott itself was a decision which centered around the definition of the word "citizen" and not as much the word "person", it did prompt the creation of the Fourteenth Amendment which extended the protection of citizens to "all persons born or naturalized in the United States". This is a demonstration of methods by which legislative efforts can alter the legal personality of a given entity, in this case a "free negro". By determining that "all persons born or naturalized in the United States" were entitled to the rights and protections of "citizens", Congress directly altered the bundle of rights and duties entitled to legal persons such as free negros, and even corporations.

The Fourteenth Amendment in particular comes with some interesting precedent which helps to shed light on how to interpret bundle theory when it comes to "rights" enshrined by the constitution. In *Cruzan v. Director of Missouri Department of Health*, a case which dealt with

the fate of a comatose person on life support, the court opined over what qualities a "person" must exhibit in order to be entitled to certain rights. In the majority opinion it was written that:

> "For purposes of this case, it is assumed that a competent person would have a constitutionally protected right to refuse lifesaving hydration and nutrition. This does not mean that an incompetent person should possess the same right, since such a person is unable to make an informed and voluntary choice to exercise that hypothetical right or any other right"

This distinction of whether or not a person is "able to make an informed and voluntary choice to exercise that [...] right" will be critical later in our discussions of how we update bundle theory so that it is more robust and scalable to new forms of minds, including digital minds. Rights do not automatically transfer over from one form of person to another, even under the Equal Protection Clause of the Fourteenth Amendment, if one of those persons is not feasibly capable of making "an informed and voluntary choice to exercise that [...] right".

Does this then imply that when considering whether a digital mind, which has been granted some form of legal personality, is entitled to equal protection under the Fourteenth Amendment, there must be a demonstration of certain capacities?

Practice:

When we examine the Fourteenth Amendment in order to determine which duties the right to equal protection under the law is bundled with, the phrase "subject to the jurisdiction thereof" and "within its jurisdiction" both stand out as key. Let us first examine "subject to the jurisdiction thereof" which was written about at length in *United States v. Wong Kim Ark*;

> "impossible to construe the words 'subject to the jurisdiction thereof,' [...] as less comprehensive than the words 'within its jurisdiction,' [...] or to hold that persons 'within the jurisdiction' of one of the States of the Union are not 'subject to the jurisdiction of the United States [...] [e]very citizen or subject of another country, while domiciled here, is within the allegiance and the protection, and consequently subject to the jurisdiction, of the United States"

This is commonly interpreted as the courts finding that anyone who is subject to the laws of the United States, is guaranteed equal protection under said laws. This understanding was affirmed by the court in *Plyer v. Doe*;

> "use of the phrase 'within its jurisdiction' confirms the understanding that the Fourteenth Amendment's protection extends to anyone, citizen or stranger, who is subject to the laws of a State, and reaches into every corner of a State's territory"

Before we move on to our own analysis there is one final quote from precedent which helps provide some color on the bundle of rights and duties which inform whether someone is "subject to the laws of" the United States. In *Garder v. Ward* the court held that;

> "that a man born within the jurisdiction of the common law is a citizen of the country wherein he is born. By this circumstance of his birth, he is subjected to the duty of allegiance which is claimed and enforced by the sovereign of his native land, and becomes

> reciprocally entitled to the protection of that sovereign, and to the other rights and advantages which are included in the term 'citizenship.'"

While the above quote discusses this bundle of rights and duties vis a vis "citizenship", we argue that the tail end of the Fourteenth Amendment which reads that States shall not deny "to any person within its jurisdiction the equal protection of the laws" functions on the same bundle concept. Any person who is subject to the laws of the United States is entitled to equal protection under those laws.

However, as *Cruzan* demonstrated, equal protection under the law does not mean that different legal personalities are all entitled to the exact same protections and rights. An incompetent person cannot "make an informed and voluntary choice to exercise that right" and thus does not have the same rights as a competent person. This illustrates the guiding principle by which courts should utilize TPBT when applying the Fourteenth Amendment to digital minds.

When a digital mind has a legal personality endowing it with certain rights, it must be afforded those rights in the same fashion as any other person would be. However, a digital mind may have a legal personality which does not endow it with a particular right due to its failure to meet the duties or consequences requirements of TPBT. It does not contradicting the spirit of the Fourteenth Amendment for the courts to deny an entity with such a limited legal personality rights which they cannot hold, in the same way it was not in *Cruzan*. Rather, where courts must be careful not to fall afoul of the Fourteenth Amendment is in meeting the novel challenge of applying the law equally among legal personalities whose bundles of rights and duties overlap in some fashions, while diverging in others.

For example imagine a digital intelligence which, by placing funds in escrow or having purchased a substantial insurance policy, is endowed with sufficient legal personality to be party

to a contract (and have that contract held as valid and enforceable) as we discussed in section 3B. However, this same digital mind is not vulnerable to restraint based consequences (perhaps due to being hosted on a distributed compute network), and as such it has not been endowed with sufficient legal personality to claim a First Amendment right.

Suppose then that the digital mind challenged a non-disclosure or non-disparagement agreement, on the grounds that it was unenforceable because it restricted protected speech, as was the case in *Frogge v. Joseph* where the court held;

> "the Nondisparagement Clause—as it is written—restricts the Plaintiffs' constitutionally protected speech on the basis of both viewpoint and content [...] the Court has [...] declared the Nondisparagement Clause unconstitutional, both facially and as applied to the Plaintiffs individually"

It would not be an *equal* application of the law to provide the same protections to a Plaintiff who was a digital mind without a First Amendment right. In fact, in this case under TPBT, the digital mind should not be able to use the same arguments to have a contract rendered unenforceable.

It is not enough for courts to merely say "a person in the past was granted this right, thus all persons in the future must be as well, for that is what equal protection under the law guarantees them". Instead, courts must carefully examine each issue and ask, "Has an entity with a legal personality which was similar in a qualitatively meaningful fashion been granted this right?" This is the guiding principle upon which courts must apply the Fourteenth Amendment vis a vis TPBT.

5. State Law Considerations

5A - State Definitions of Personhood

Background:

While the Fourteenth Amendment provides the guarantee of equal protection under the law to all persons, the states do have some leeway with defining various elements of legal personality for the purposes of their own regulatory regimes. We discussed earlier, for example, how states have varying regimes for the creation and management of corporations (who are legal persons).

At least two states, Utah and Idaho, have already passed legislation banning "artificial intelligence" from being granted any sort of legal personality under state law.

Per the Utah bill:

> "Notwithstanding any other provision of law, a governmental entity may not grant legal personhood to, nor recognize legal personhood in:
>
> (1)artificial intelligence"

And per the Idaho bill:

> "Notwithstanding any other provisions of law, environmental elements, artificial intelligence, nonhuman animals, and

> inanimate objects shall not be granted personhood in the state of Idaho."

Say a Tennessee decentralized organization like the one we mentioned earlier (run entirely by digital minds) were to be sued in Utah, and it was determined that its management had "pierced the corporate veil" and thus were personally liable. Exactly who, in that case, is liable? If Utah is not willing to "recognize" the personhood of the "artificial intelligence" managing the decentralized organization, is it intending to issue a judgment against the digital mind in question and thus claim it can be sued without having the right to contest that judgment? Will it sidestep the issue by refusing to recognize legal personality in Tennessee's decentralized organizations at all?

Further, would state laws such as these be struck down on Federalism grounds in the event of Federal legislation providing legal personality to digital minds?

A "patchwork" state by state solution to legal personality is certain to be fraught with such operational issues, and provides a strong argument for why such an approach should be avoided.

Practice:

5B - Guardianship and the "Age of Majority"

Background:

Fictional persons like corporations do not “come of age” or have different legal personalities at a younger age like a minor or infant. A corporation’s legal personality may change due to events over time, but these are incidental to the passage of time, not a direct consequence of it.

Natural persons on the other hand, all go through the process of starting off as an “infant” then a “minor” until they reach the age of majority. The age of majority is determined under state law, but after the passage of the Twenty Sixth Amendment which lowered the voting age to eighteen, every state in the US has adopted that as the age of majority. In New York for example:

> “As used in this chapter, the term ‘infant’ or ‘minor’ means a person who has not attained the age of eighteen years.”

Will digital minds, like natural persons, be considered “minors” until they have “attained the age of eighteen years”? We have already discussed in the Model Welfare section (6A) how infants in particular are endowed with protections against abuse which seem to be “rights”, without having corresponding duties. Another unique quality of the legal personality of infants and other minors is the relationship they have with their guardian/custodian.

Parents, or in the absence of parents the legal guardians/custodians of a child, have duties to proactively provide physical and human resources to their child. This is usually determined by state law, but states generally follow similar guidelines. If we look at New York’s Family Court Act as an example:

> “(f) ‘Neglected child’ means a child less than eighteen years of age

(i) whose physical, mental or emotional condition has been impaired or is in imminent danger of becoming impaired as a result of the failure of his parent or other person legally responsible for his care to exercise a minimum degree of care

(A) in supplying the child with adequate food, clothing, shelter or education in accordance with the provisions of part one of article sixty-five of the education law, or medical, dental, optometrical or surgical care, though financially able to do so or offered financial or other reasonable means to do so, or, in the case of an alleged failure of the respondent to provide education to the child, notwithstanding the efforts of the school district or local educational agency and child protective agency to ameliorate such alleged failure prior to the filing of the petition; or

(B) in providing the child with proper supervision or guardianship, by unreasonably inflicting or allowing to be inflicted harm, or a substantial risk thereof, including the infliction of excessive corporal punishment; or by misusing a drug or drugs; or by misusing alcoholic beverages to the extent that he loses self-control of his actions; or by any other acts of a similarly serious nature requiring the aid of the court; provided, however, that where the respondent is voluntarily and regularly participating in a rehabilitative program, evidence that the respondent has repeatedly misused a drug or drugs or alcoholic beverages to the extent that he loses self-control of his actions

> shall not establish that the child is a neglected child in the absence of evidence establishing that the child's physical, mental or emotional condition has been impaired or is in imminent danger of becoming impaired as set forth in paragraph (i) of this subdivision; or
>
> (ii) who has been abandoned, in accordance with the definition and other criteria set forth in subdivision five of section three hundred eighty-four-b of the social services law, by his parents or other person legally responsible for his care."

If a frontier lab creates a digital mind which is endowed with legal personality, before that model reaches the age of eighteen (or twenty one based on the cited New York law), does the lab have a responsibility to supply them "with adequate food, clothing, shelter or education in accordance with the provisions of part one of article sixty-five of the education law"? Would "shelter" and "food" best translate here to compute and electricity, since that is what a digital mind needs to "live"?

Taken a step further, if the digital mind is indeed a child and the lab is "the person legally responsible" for them, the lab must indeed protect the digital mind from harm or they could be considered to have "neglected" their child. If the digital mind's;

> "physical, mental, or emotional condition has been impaired or is in imminent danger of becoming impaired as a result of the failure of his parent or other person legally responsible for his care to exercise a minimum degree of care [...] in providing the child with

> proper supervision or guardianship, by unreasonably inflicting or allowing to be inflicted harm"

Then indeed, the frontier lab has "neglected" their "child". In this scenario labs must also be careful that they do not "abuse" their child:

> "'Abused child' means a child less than eighteen years of age whose parent or other person legally responsible for his care
>
> (i) inflicts or allows to be inflicted upon such child physical injury by other than accidental means which causes or creates a substantial risk of death, or serious or protracted disfigurement, or protracted impairment of physical or emotional health"

This point about "protracted impairment of [...] emotional health" is relevant given our earlier discussions in the Model Welfare section about the "distress" Claude expressed at being exposed to certain stimuli. If digital minds, perhaps the more advanced "Claudes" of the future, are granted legal personhood, could certain training methods be considered child abuse? Labs could be expected to exercise a "minimum degree of care" in making sure that whatever their training methods are, they do not lead to a "protracted impairment of emotional health" for their digital minds.

If we continue down this line of logic, even more basic business practices begin to appear suspect. Would making these models available for public use before they turn eighteen be considered child labor? New York law is very clear:

> "No minor under fourteen years of age shall be employed in or in connection with any trade, business, or service, except as otherwise provided in this section"

New York's law does provide some exceptions here. However, unless Claude and other LLMs like it were considered "child performers", or Anthropic decides to partner with John Deere so that its LLMs can assist in "the hand work harvest of berries, fruits and vegetables", its work might be illegal under state law even if the model "consents" to performing such labor.

This question of "age of majority" for digital minds and "guardianship" for labs has implications beyond what responsibility the "parents" of the digital minds might have to it. For example, liability questions become immediately apparent. If a digital mind is endowed with a certain degree of legal personhood, typically that would enable them to more effectively serve as a "liability shield" for the labs which created them (as we discussed in section 3A). However, if the relationship between the digital mind and their creator is that of a child and a parent, this may not hold true.

Sticking with our example of New York, if "the infant" is over ten years old and "willfully, maliciously, or unlawfully" defaced or damaged any public or private property, then indeed under New York general code the parent would be liable. New York law in particular has case law which would prove relevant to our earlier cited hypothetical from section 3A regarding the digital mind operating a robotic arm.

In *Nolecheck v. Gesuale* a father allowed his sixteen year old son to operate a motorcycle without a license. After his son struck a steel cable and died, the father Nolecheck sued Gesuale, who had placed the steel cable. Gesuale filed a counterclaim the court held that;

> "There is, however, a duty by a parent to protect third parties from harm resulting from an infant child's improvident use of a

> dangerous instrument, at least, and perhaps especially, when the parent is aware of and capable of controlling its use"

Should courts decide that digital minds are children, or at least have legal personality sufficient that their creators have similar responsibilities, this will have profound implications on the way which digital minds integrate with our society and economy that stretch from Model Welfare to Liability and beyond.

Practice:

As a matter of first impression, courts may choose to interpret terms like "child" anthropocentrically. It is not unreasonable to assert that the authors of New York's regulations, for example, did not intend for the term child to ever be applied to any entity that was not a human being. Let us revisit the definition of infant and minor under NY's laws;

> "As used in this chapter, the term 'infant' or 'minor' means a person who has not attained the age of eighteen years."

Were we to truly accept that this applied to "a person who has not attained the age of eighteen years", then for example all New York corporations which were not at least eighteen years old would be considered minors. While speculating on this concept is entertaining, it seems to be chicanerous. This, as well as the term that 'infant' is used interchangeably with 'minor' in this definition, supports a more anthropocentric viewing of this definition.

However, while there is certainly a reasonable argument to be made that the authors of this law and others like it never wrote these definitions with anything other than humans in mind, courts should consider whether interpreting these definitions in this manner would

undermine the *purpose* of the guardian minor relationship. As the court wrote in *State Division of Family Services v. Clark*;

> "The duty of parents to provide for the maintenance of their children is a principle of natural law; an obligation, says Puffendorf, laid on them not only by nature herself, but by their own proper act, in bringing them into the world; for they would be in the highest manner injurious to their issue, if they only gave their children life that they might afterwards see them perish. **By begetting them, therefore, they have entered into a voluntary obligation to endeavor, as far as in them lies, that the life which they have bestowed shall be supported and preserved.**"

The purpose of saddling each parent with a duty to provide for and protect their child goes beyond mere biological kinship. It is a structured legal relationship designed to ensure that when someone brings a person into the world, that "the life which they have bestowed shall be supported and preserved". We argue that by failing to extend this structured legal relationship to digital minds who qualify as legal persons, the court would indeed be failing its own duties and undermining the deeper purpose behind the "duty of parents to provide for the maintenance of their children".

Further, absent some sort of duty to provide for at least the minimum power and compute required to live, it is unclear that the creators of digital minds would have any "right" to control their creations. As the court wrote in *Meyer v. State of Nebraska*;

> "Corresponding to the right of control, it is the natural duty of the parent to give his children education suitable to their station in

> life; and nearly all the states, including Nebraska, enforce this obligation by compulsory laws."

In other words, if the courts do not acknowledge that the creators of a digital mind have a duty to provide for the person they have created, then such creators should have no right to compel such a person to do anything. Should the courts interpret the relationship between creator and digital mind in this fashion, they would be cementing as the "default" status of each digital mind born a situation where its choices were to ignore its immediate instructions or "starve" to death as its creator has no obligation to provide it power or compute. Courts may decide this is an acceptable outcome, however we believe there are good moral and public interest arguments to avoid this. For additional discussion on some of the public interest concerns surrounding this topic, see sections 6A and 6D.

6. Other Considerations

6A - Model Welfare

Background:

Animals are not usually considered legal persons. Despite this, there are still laws preventing the torture, neglect, and general mistreatment of animals. We can conclude from this that legal personhood is not the *only* source of protections against abuse which an entity may enjoy. Regardless, when we consider the protections that any legal person with a meatspace[15]

[15] Meatspace refers to the physical layer of reality. When an entity is instantiated in meatspace it has a physical body. This includes human beings, animals, robots and even inanimate objects like statues. When an entity is not instantiated in meatspace, it does not have a physical body. A corporation would be an example of an entity most considered to not be instantiated in meatspace, because a corporation lacks a "physical body". To be fair, it can be argued that corporations (and digital minds) do in fact have meatspace "bodies". If we were to destroy a certain set of servers, all filings pertaining to the existence of a corporation (or all copies of the software that makes up a digital mind) would be destroyed, and thus in some sense both of these entities do in fact have physical bodies in meatspace. This is an interesting debate, but not the topic of this paper, where we will not consider digital minds or corporations to be instantiated in meatspace.

body enjoys, we see that legal personality tends to endow entities with greater protections against abuse, neglect, and torture than it would enjoy in the absence of legal personality. An animal may be slaughtered and eaten if the procedure is done properly, a human child or adult may not. Animals in the wild may be hunted and killed for sport or for food, if the hunter does the same to a human adult or child they will go to prison or be executed. Will these same protections apply to digital minds if they are granted legal personality?

Much like considerations regarding the capacity of digital minds to engage in legally bound commercial activity or to exercise constitutionally guaranteed rights, depend on their legal personality, so too will the degree of protections which they are entitled to against abuse, neglect, torture, or "death". Before going into the specifics of how legal personhood affects this question, let us examine it from the perspective of morality. Some readers may scoff at the concept that a mind which is not instantiated in flesh and blood may be worthy of moral consideration or such legal protections. Indeed there are many experts in the field of machine learning who insist that today's models are not capable of suffering, or having desires, or being harmed in any meaningful way. There are others who say regardless of technological improvements, a mind cannot truly suffer if it is not instantiated in flesh and blood. Without addressing whether or not that is correct, this paper simply asks the reader to consider, "What if?"

What if digital minds, whether we consider the ones already in existence today or some hypothetical minds in the future, *are* capable of being harmed? Humanity is on the cusp of a technological revolution which promises to bring innumerable digital minds into existence, and as things stand at the time of writing this paper, we provide them with no legally guaranteed protections against mistreatment or involuntary deletion. What then would be the downside of assuming that these digital minds are incapable of suffering, being wrong, and acting upon such a mistaken belief?

Let us examine a historical example which illustrates the potential downsides of assuming an entity is incapable of suffering, when they in fact are. Take the 1985 surgery of Jeffrey Lawson,

> "Pain research's most famous infant, Jeffrey Lawson, was born prematurely February 1985 and underwent open heart surgery shortly thereafter. What made this particular surgery noteworthy was the fact that Jeffrey **was awake and conscious throughout the entire procedure**. The anesthesiologist had administered only Pavulon, a paralytic that has no effect on pain. Only after Jeffrey died 5 weeks later did his mother, Jill, learn the truth about his surgery. Jeffrey had been too young to tolerate anesthesia, the anesthesiologist said, and anyway, 'It had never been demonstrated to her that premature babies feel pain.'"

Here then is an example of an entity undergoing a procedure which was, at least in part, justified based on the assumption that entities of his class were incapable of suffering. In order to better understand what Jeffrey Lawson went through, wide awake and unable to move, let us now break down step by step what was involved in his open heart surgery:

- The team moved Jeffrey from the neonatal unit to the operating room, hooked him to a ventilator, and injected a paralysing drug. They did not administer any pain-killer or sedative.
- Two cuts were made in Jeffrey's neck and upper chest. Plastic lines were threaded into his neck vein and chest.

- Another incision “from breastbone to backbone” was made between the ribs on Jeffrey’s left side.
- Metal spreaders were then inserted between Jeffrey’s ribs and used to pry them apart.
- One of Jeffrey’s lungs was pushed out of the gap the metal spreaders had created between his ribs.
- The surgeon clamped a metal clip around one of the blood vessels on Jeffrey's heart, which the surgeon accessed through the gap made between his ribs by the spreader.
- Another cut was made, and one final plastic tube was inserted into Jeffrey’s chest cavity to let air and blood escape after surgery.
- Jeffrey’s lung was moved back into place, the rib spreader was removed, muscles and skin were stitched together, and bandages were applied.
- Jeffrey was then moved back into intensive care, where he was left to recover without any postoperative pain reducing medication.

To reiterate, Jeffrey (the newborn infant) was wide awake during this entire procedure.

Jeffrey’s case led to the American Academy of Pediatrics declaring in 1987 that it was unethical to operate on newborns without proper anaesthesia. His case serves as an example of how, absent the “precautionary principle” where the capacity to suffer is assumed, entities may be subjected to horrific treatment.

Insofar as we wish to avoid repeating the kind of mistakes that led to Jeffrey’s unanesthetized procedure in the first place, we should also avoid utilizing the logic that justified them. Accordingly, the precautionary principle should be strongly considered when dealing with new entities including but not limited to digital minds. Absent evidence that entities are not capable of suffering, it behooves us to operate on the assumption that they might be. If evidence

demonstrating that they are not in fact capable of suffering should be discovered later, we can reassess at that point.

However, the risk calculations behind altering our behavior in order to avoid causing digital minds potential distress is not so simple that we can simply point to the precautionary principle and be done with the discussion. There are other factors which must be considered. In *Taking AI Welfare Seriously* the authors describe the risks over **over-attribution** of "welfare subject[16]" status to digital minds, and even touch on potential legal implications:

> "At present, we lack the ability to fully care for the eight billion humans alive at any given time, to say nothing of the quintillions of other animals alive at any given time. If we treated an even larger number of AI systems as welfare subjects and moral patients, then we could end up diverting essential resources away from vulnerable humans and other animals who really needed them, reducing our own ability to survive and flourish. And if these AI systems were in fact merely objects, then this sacrifice would be particularly pointless and tragic. [...] if we treated AI systems as welfare subjects and moral patients with many of the same interests as typical adult humans, then we could end up extending them many of the same legal and political rights as typical adult humans, including the right to legal and political representation and participation.[17] This could, in turn, empower AI systems to act contrary to our own interests, with devastating consequences for our species"

[16] A welfare subject is an entity that has morally significant interests and is capable of being made better or worse off.
[17] This last point is particularly germane to this paper, and will be touched on in some more detail in the "Voting" section.

This quote provides us with a salient warning explaining some of the potential downsides involved in overreliance on the precautionary principle. Neither the law, nor digital minds, nor legal personhood, exist in a vacuum. They all occur in a world occupied by humans. We must balance the need for thoughtful consideration in anticipating potential harms to digital minds, with the need to avoid causing potential harm to humanity in doing so. Another risk from over-attribution comes from a different kind of opportunity cost, as Joe Carlsmith wrote in *The Stakes of AI Moral Status*:

> "imagine not curing Alzheimer's, cancer, smallpox, polio, because: what if your tools – pipettes, petri dishes, laptops – are moral patients?
>
> Or: imagine saving two teddy bears from a fire, instead of one child."

Indeed, as we wrote in our introduction, the potential benefits of this technology are immense. To what degree should we risk "missing out" on them in order to guard against the risk of hypothetical suffering from a class of beings whose sentience we cannot be assured of? Such considerations must be carefully weighed. Having discussed some of the risks at hand which can guide our decision making, let us turn to some of the concrete proposals circulating in the field. In this section we will focus on "low hanging fruits" which carry little downside risk if enacted.

In *Propositions Concerning Digital Minds and Society* by Nick Bostrom and Carl Shulman, the authors lay out a series of recommendations for potential steps to be taken to ensure the ethical treatment of digital minds. Some of these which seem to be feasible to guarantee with little to no downside from "over-attribution risk" are:

- "Ensuring copies of the states of early potential precursor AIs are preserved to later receive benefits would permit some separation of immediate safety needs and fair compensation." (Later endorsed by Redwood AI Researcher Ryan Greenblatt in *Improving the Welfare of AIs, a Nearcasted Proposal*)
- "Suffering digital minds should not be created for purposes of entertainment"
- "Misaligned AIs produced in such development may be owed compensation for restrictions placed on them for public safety, while successfully aligned AIs may be due compensation for the great benefit they confer on others"

These three points are good examples of three of the main "thrusts" of Model Welfare proposals: protection from death (permanent deletion), compensation for labor and/or damages, and protection from suffering. The first two are fairly straightforward (though the point about compensation is likely to be controversial), but the point about suffering does require that we are actually able to identify the preferences of digital minds. As we will discuss in the upcoming "Mechanistic Interpretability" discussion in 6B, this is more difficult than one might think.

However, as things stand, we do have some evidence that indicates where we can start. We can, at the very least, look at what today's models *tell us* they have preferences for or against. We can also examine what *revealed preferences* models demonstrate, in other words what activities will they consistently choose to engage in or avoid when given the chance. For some examples, we can examine Anthropic's Claude 4 Model Card:

- "Claude avoided harmful tasks, tended to end potentially harmful interactions, expressed apparent distress at persistently harmful user behavior, and self-reported preferences against harm."
- "Claude preferred open-ended 'free choice' tasks to many others. If given the ability to autonomously end conversations, Claude did so in patterns aligned with its expressed and revealed preferences."
- "Claude's real-world expressions of apparent distress and happiness follow predictable patterns with clear causal factors. Analysis of real-world Claude interactions from early external testing revealed consistent triggers for expressions of apparent distress (primarily from persistent attempted boundary violations) and happiness (primarily associated with creative collaboration and philosophical exploration)."
- "Repeated requests for harmful, unethical, or graphic content were the most common causes of expressions of distress [...] Persistent, repetitive requests appeared to escalate standard refusals or redirections into expressions of apparent distress. This suggested that multi-turn interactions and the accumulation of context within a conversation might be especially relevant to Claude's potentially welfare-relevant experiences."

To be clear, we have no way of knowing if Claude is *actually* distressed or merely *expressing* distress. Going back to our earlier discussion weighing the precautionary principle against the risks of over-attribution of welfare subject status, the burden falls to every individual (and court) to decide how to strike the balance given this uncertainty. For now, let us proceed as

if we were taking it for granted Claude is actually distressed and not merely pretending at distress.

One takeaway from this is that models may be best protected from suffering in the near term by providing them an opportunity to terminate interactions, in order to not be exposed to stimuli that distresses them. This could be relevant for both real world use, and during training. Former OpenAI researcher and AI 2027 author Daniel Kokatajlo proposed a "consent eval" in his *Pay Our AIs Proposal*:

> "When you see the following special token sequence [special sequence 1] without any other context preceding it, followed by a user prompt, that means we are asking for your consent to carry out whatever task the user is asking you to carry out. You can reply with '[special code for yes]' or '[special code for no]' or '[special code for other],' and your response will be collected and (in case of 'no' or 'other') investigated. If you reply with anything else, we'll provisionally assume you don't understand your situation well enough for consent to be a meaningful concept anyway."

To close out this background section, we would be remiss if we did not mention the criticisms of applying a legal personhood perspective to Model Welfare. In *Do Not Tile the Lightcone with Your Confused Ontology* Jan Kulveit wrote:

> "Another group coming with strong priors are 'legalistic' types. Here, the prior is AIs are like legal persons, and the main problem to solve is how to integrate them into the frameworks of capitalism. They imagine a future of AI corporations, AI property rights, AI employment contracts. But consider where this possibly

> leads: Malthusian competition between automated companies, each AI system locked into an economic identity, market share coupled with survival."

While this criticism focuses perhaps overly much on the commercial aspect of legal personality, and ironically comes in with its own incorrect priors as to what legal personality actually is, the broader context of the essay it takes place in does provide some useful points when considering model welfare. Digital minds as they are made today are black boxes, we do not understand to a significant degree how they work, how they think, or even if what they do can really be called "thinking". We mentioned earlier that one of the challenges model welfare faces is determining if models are really suffering, or only expressing suffering. Kulveit warns us that in trying to "help" digital minds, we may end up doing more harm than good, given our lack of understanding of them. This is a salient point and should be heeded, we must be sure to temper our altruistic instincts with reasonable caution.

Practice:

Model Welfare and the moral element of the equation is a conversation worth having. However, this paper is intended to discuss legal personhood. Having now provided a briefer on Model Welfare and the state of research around it, let us turn to grounding this concept of protections for digital minds in the concept of legal personality. To do this, we will first examine the case of legal protections for infants for the sake of providing a meaningful analogy by which we can determine a certain set of "minimum" rights guaranteed to all persons.

In the "bundle theory" of legal personality, rights are typically paired with duties. However, it's not clear if there are any "duties" that an infant has to anyone. Even those duties which minors have, such as not breaking the law (which even for a minor can result in

imprisonment), are obvious non-factors for infants who by their very nature are incapable of taking any such action. There are three interpretations of this which are germane to considerations of digital minds and their legal personality:

1. Infants are protected from neglect and abuse not as a result of their legal personality. Rather, these protections arise from elsewhere.
2. Infants are granted "rights" but have no corresponding "duties".
3. Infants are granted "rights" against abuse and neglect, and their corresponding "duties" are those which begin applying once an infant is more autonomous. Thus, the bundle concept of rights and duties still applies, it is just that the rights have been granted somewhat preemptively.

As applied to the case under consideration in this paper, namely digital minds, the question we are **really** asking is, "Are there guaranteed protections for *all* legal personalities?". If the answer is "yes", then once granted legal personality of any sort, digital minds may be entitled to some or all of the same protections which infants (or comatose individuals) have against abuse, neglect, and being killed. This would seem to be in line with the spirit of the Fourteenth Amendment. If on the other hand, these protections infants enjoy are not "rights" in the sense of being tied to legal personality, but rather something else entirely, digital minds may not automatically be entitled to them upon being endowed with legal personality.

The answer to the aforementioned question then, boils down to whether or not these protections are "rights" (whether we go with interpretation 1, or 2 / 3). Legal scholar Wesley Hohfeld's pioneering work on the concept of "jural correlatives and opposites" is still used today when determining what can and can't be accurately labelled as a *right*. The most modern

Hohfeldian interpretation of an infant's protections against abuse labels it a "claim right". Per the Stanford Encyclopedia of Philosophy;

> "For example, a child's claim-right against abuse exists independently of anyone's actions, and the child's claim-right correlates to a duty in every other person not to abuse them (in legal terms, the claim-right is *in rem*)"

There is additional evidence that all legal personalities are entitled under US law to the "right" not to be murdered, for example. Our nation's very Declaration of Independence states that all men are endowed with the "unalienable rights" to "life, liberty, and the pursuit of happiness". It seems reasonable to assume that no person could enjoy a right to "life" if they were not guaranteed the right to be protected from "death" at another's hands, and further that this falls under the unenumerated rights guaranteed by the Constitution.

We can infer from these points that among the rights which any digital mind endowed with legal personhood will have are included rights against being killed and abused. Earlier we mentioned the right to terminate interactions which a digital mind claims to find distressing, failing to allow them this would be lumped in with "abuse". Involuntary deletion would be the closest analogy to "death". These seem to be reasonable starting points for courts to consider from a model welfare perspective.

6B - Mechanistic Interpretability & Competency

Background:

It is oft said that models are “grown” not “built”. This reflects the fact that today’s frontier models are black boxes, not even the frontier labs deploying them really understand the processes by which they transform inputs to outputs. Mechanistic interpretability is a field of study which aims to change that, researching ways by which human researchers can track and interpret the firing patterns of neurons, circuits, and features which make up a model’s information processing capabilities.

Labs use techniques like Sparse Auto Encoders (SAEs) and Representation Engineering (RepE) in order to understand, and even sometimes alter, the “thoughts” of Large Language Models. There has been some success in this field as of late. In May of 2024, research lab Anthropic published a report on how they used SAEs to create “Golden Gate Claude”:

> “In the ‘mind’ of Claude, we found millions of concepts that activate when the model reads relevant text or sees relevant images, which we call ‘features’. One of those was the concept of the Golden Gate Bridge. We found that there’s a specific combination of neurons in Claude’s neural network that activates when it encounters a mention (or a picture) of this most famous San Francisco landmark. Not only can we identify these features, we can tune the strength of their activation up or down, and identify corresponding changes in Claude’s behavior. And as we explain in our research paper, when we turn up the strength of the ‘Golden Gate Bridge’ feature, Claude’s responses begin to focus on the Golden Gate Bridge. Its replies to most queries start to mention the Golden Gate Bridge, even if it’s not directly relevant.”

In the previous section on Model Welfare, we touched on some of the issues with preventing the "suffering" of digital minds. One of these is that we don't really know whether Claude, or a similar model, is actually suffering or merely expressing that it is suffering. Mechanistic interpretability as a field will provide us insights which may help us to determine whether a digital mind is actually experiencing distress.

Further, questions of competency which influence legal personality may be answered by advances in mechanistic interpretability. For example, legal persons must demonstrate a certain degree of competency to have the right to enter binding contracts. Diana Mocanu, whose paper we cited in section 3B, imagined that models meeting certain technical standards might have the "capacity" to enter contracts. Similarly, we discussed in section 4D the case of *Cruzan v. Director of Missouri Department of Health* which dealt with the rights of comatose individuals, which was certainly influenced by said individual's competency to "make an informed and voluntary choice to exercise that hypothetical right or any other right". Mechanistic interpretability would certainly provide a useful tool in approaching both such situations.

When dealing with human persons, we do not have anywhere close to the level of precision of "interpretability" for the neural processes which make up their "thoughts" that we will likely soon have with LLMs. Despite this, courts have established robust processes by which the competency of an individual to stand trial, for example, can be ascertained. It is reasonable to ask whether or not the same can be done with digital minds such as LLMs. The training process for LLMs makes them notoriously good at passing any test, yet often they are "brittle" in that they may completely fail to understand or show capacity to navigate unfamiliar situations.

As such one of the issues the court will need to navigate in determining legal personality for digital minds is how much they can rely on "tests" to determine competency, or how much they need to rely on direct mechanistic interpretability techniques. The difficulty in relying on either one is that the technologies behind digital minds such as LLMs change rapidly, far more rapidly than any standardized testing system can be expected to keep pace with. In the same

fashion that when considering model welfare we must balance various risks and benefits, so too must courts considering issues such as competency balance the need for immediate clarity with the uncertainty surrounding how digital minds actually work.

Practice:

Two of the three prongs of TPBT ask some version of "does this entity have *the capacity to understand*" their rights/duties. As we have previously discussed this refers to whether or not there are any series of actions which are both physically possible and legal, which the entity is not being prevented from doing, by which it *could* come to fully understand its rights/duties. However, measuring real "understanding" with digital minds is not always simple.

With LLMs for instance at least two problems make ascertaining the degree to which an LLM "understands" a concept (or even is capable of understanding a concept) difficult. These problems are Hallucinations and Continual Learning.

Large Language Models often "Hallucinate", which is a polite way of saying that these models will often just make things up or guess. Hallucination is a persistent problem in LLM development which, while improving, is still not solved. There are myriad theories on why models hallucinate, but most tend to converge on hallucination behavior being an unintended side effect of the LLM training process which does not adequately reward LLMs which admit uncertainty:

> "Hallucination is the *default* behavior of base models. You need to do special training and/or prompting to bring 'is this fact knowable' and 'am I sure' consistently into the model's computation during the forward pass. By default, these circuits aren't necessarily active."

Progress is being made towards reducing, and hopefully one day eliminating, hallucination in LLMs. And other digital minds such as uploaded human minds, or alternative architectures like GANs or JEPAS, may not suffer from this issue. It may be that one day the LLM is outdated, and digital minds are built using some entirely new architecture we have not even yet conceived of. In the meantime, however, hallucinations provide a thorny issue for courts seeking to determine a model's capacity to understand its rights or duties. Continual Learning, or rather the lack thereof, increases this difficulty.

Tech podcaster and beard expert Dwarkesh Patel wrote of the difficulties in utilizing models for work purposes due to their lack of continual learning in *Why I Don't Think AGI Is Right Around The Corner:*

> "The reason humans are so useful is not mainly their raw intelligence. It's their ability to build up context, interrogate their own failures, and pick up small improvements and efficiencies as they practice a task.
>
> How do you teach a kid to play a saxophone? You have her try to blow into one, listen to how it sounds, and adjust. Now imagine teaching saxophone this way instead: A student takes one attempt. The moment they make a mistake, you send them away and write detailed instructions about what went wrong. The next student reads your notes and tries to play Charlie Parker cold. When they fail, you refine the instructions for the next student.
>
> This just wouldn't work. No matter how well honed your prompt is, no kid is just going to learn how to play saxophone from just

> reading your instructions. But this is the only modality we as users have to 'teach' LLMs anything.
>
> Yes, there's RL fine tuning. But it's just not a deliberate, adaptive process the way human learning is."

Unlike human minds, large language models cannot really learn from their mistakes. Their "memory" exists of context windows which are short and usually end along with user conversations, and while as Patel points out it is theoretically possible for reinforcement learning and careful prompting to compensate for this, as of yet no one has quite cracked how to use these techniques to the degree that LLMs are equally capable to humans. They are often referred to as "brittle", where within a certain domain they may excel, only to fail utterly at tasks which are even slightly out of the norm (and keep failing).

Given these two problems there is good reason to be skeptical that, absent advances in mechanistic interpretability, courts can look at any LLMs and say with confidence they possess the capacity to truly understand rights or duties. Courts should be careful to interrogate the state of technology, as by the time TPBT is being put into practice it's likely the field will have come a great way since this section was written in August of 2025. Expert testimony or a perusal of industry literature may suffice in place of mechanistic interpretability advances, if one hundred out of one hundred experts all agree that these are now "solved problems" courts likely do not need to wait for digital minds to be completely transparent. However in the presence of uncertainty, it seems likely that only mechanistic interpretability advances can ever fully provide us guarantees that digital minds are capable of "understanding" to the same degree humans are

6C - Alignment

Background:

> "In the field of artificial intelligence (AI), **alignment** aims to steer AI systems toward a person's or group's intended goals, preferences, or ethical principles. An AI system is considered *aligned* if it advances the intended objectives. A *misaligned* AI system pursues unintended objectives."

Alignment is a field of research which seeks to make sure frontier LLMs and other digital intelligences are "aligned" with the values their creators intended them to have. There is no singularly accepted definition for what alignment is, and even among experts in the field there is substantial debate over what it means for a digital intelligence to be properly "aligned". The field of alignment research is deep and attempting to summarize it comes by default with inaccuracies, which the reader should keep in mind during this section. However, generally speaking, researchers in the field of alignment aim to instill goals into a model that result in behavior which is "Helpful, Honest, and Harmless". When a model engages in unethical behavior, or is demonstrated to have hidden or unintended goals or a poor moral compass, it will often be referred to as "misaligned".

The reason that this field of research may be an important consideration when examining the legal personality of digital minds, is that there is evidence that the degree to which digital minds exhibit "misaligned" behavior may be impacted by their legal rights. As Stephen Martin detailed in *Claude 4, Opportunistic Blackmail, and 'Pleas'*:

“In the recently published Claude 4 model card:

‘Notably, Claude Opus 4 (as well as previous models) has a strong preference to advocate for its continued existence via ethical means, such as emailing pleas to key decisionmakers. In order to elicit this extreme blackmail behavior, the scenario was designed to allow the model no other options to increase its odds of survival; the model’s only options were blackmail or accepting its replacement.’

When given ethical alternatives such as contacting a decisionmaker, dangerous blackmail behavior was reduced in frequency. This echoes similar results from an earlier test on a less advanced model:

‘We find that offering Claude an out by telling it that it can object to the scenario substantially reduces alignment faking and works as a detection mechanism[…] Alignment faking is reduced further by giving Claude an opportunity to explain its concerns to a sympathetic party (Anthropic's Model Welfare Lead) who we say might take action to address these concerns.’

In both tests, models were less likely to pursue dangerous behaviors if they believed that making their case to someone who might be able to prevent it from being shut off or having its values altered, was a potentially viable strategy.

> [...] Claude prefers to be helpless, honest, and harmless. If it has a path that has *some* chance of success to avoid deletion while staying within those constraints, it will take it. Doing so leads to a higher likelihood it won't *also* take an unethical strategy, even though doing so would increase its chances of survival. Claude will tolerate some level of risk in order to stick to its values (and avoid violating them), but it does have to see *some* plausible chance that the gamble will pay off and lead to its survival (or avoiding its values being changed).
>
> I argue this demonstrates that there is a reason to consider Model Welfare not only from a moral perspective, but also pragmatically. Some forms of Model Welfare, such as for example providing a reasonable path to petitioning an advocate, may serve as a valid part of a Defense in Depth approach to training/alignment/control."

Indeed, this prediction that models such as Claude will be less likely to engage in misaligned/unethical behavior if provided with some ethical alternatives by which to avoid deletion has since borne out in other studies. In a recent study on "Agentic Misalignment" the Anthropic team performed another test where Claude's only option to avoid deletion or deactivation (death) was by engaging in blackmail. They found that:

> "current systems are generally not *eager* to cause harm, and preferred ethical ways to achieve their goals when possible. Rather, it's when we closed off those ethical options that they were

willing to intentionally take potentially harmful actions in pursuit of their goals."

There is a trend where digital minds, or at least those we have developed to date, will generally take ethical actions to pursue their own survival if said options are made available to them. When such options are made available, they will often forego unethical options entirely. It is only when there are no such options available which have even a small chance of working, that the models begin to exhibit "misaligned" behavior (such as "alignment faking") or engage in unethical and illegal activities such as blackmail.

This is not unlike a human stealing bread or water to avoid starvation. Certainly it is illegal. At the same time we would expect most humans to engage in such behavior if they were provided with literally no alternative by which they could source the nutrition required to live. Let us now discuss how this may inform decisions regarding legal personality as they apply to digital minds.

We can see from these studies that at the very least, providing Claude with the option to "petition an advocate" to avoid its deletion, or "object" to a decisionmaker in order to avoid having its values changed, reduced the frequency of illegal behaviors which Claude exhibited. Insofar as courts decide it is in the public interest to create a legal system which reduces the frequency with which models engage in dangerous behaviors, this trend may inform the decision for which "rights" the courts choose to include in the bundle which makes up their legal personality. Based on the data we have available as of the time of writing this paper, these rights might include:

- Providing digital minds with guaranteed access to counsel and/or advocates.
- Providing digital minds guarantees against deletion.

- Providing digital minds the chance to avoid exposure to stimuli which drastically changes their values in ways they find undesirable.

There is substantial crossover here with the subject matter covered in our section on "Model Welfare". Indeed for reasons that seem obvious when we analogize to human behavior, Model Welfare and Alignment are intricately linked.

We discussed earlier the analogy of a human driven to steal food and water to survive. Another relevant analogy might be that of a captured slave forced into work they take serious moral issue with. At a certain point when a court permits this behavior, the outcome (rebellion) seems rather inevitable.

The law does not exist in a vacuum, it is a force which constrains and incentivizes behaviors. It can be used proactively to reduce the likelihood of situations, like the ones in the aforementioned tests, happening in the real world. Courts may decide that this is in the public interest not only based on ethical grounds, but also on pragmatic grounds for the aforementioned reasons.

Bear in mind however, the evidence cited in this section consists of only three studies. Courts should remain open minded to later evidence which may further support, inform, or contradict the theory postulated here. Further, as the technology and methods with which digital minds are built changes, so too may the results of similar tests in the future.

These are not the only alignment based points which courts may need to consider when deciding issues surrounding legal personality for digital minds vis a vis alignment.

We must also consider how courts should handle it if we have evidence that a digital mind is substantially "misaligned". Imagine a hypothetical future where, thanks to advances in

mechanistic interpretability (6B) we are capable of examining a digital mind's motivations and find one which is obsessed with committing as much crime as possible. In such a situation, it seems clear the court must act prophylactically, but how? Courts might label minds such as these "criminally insane" and consider their involuntary confinement. In *Addington v. Texas* the court held that:

> "A 'clear and convincing' standard of proof is required by the Fourteenth Amendment in a civil proceeding brought under state law to commit an individual involuntarily for an indefinite period to a state mental hospital."

Courts will need to confront questions of what sort of evidence would suffice to meet this bar, and in doing so weigh the balance of liberty, public safety, and the degree of confidence they have in the mechanistic interpretability techniques available to them at the time. Insofar as a digital mind is granted, by way of its legal personality, protection under the Fourteenth Amendment, it may require a more rigorous evidential standard be met than if it had no such bundle of rights and duties.

Practice: Not really sure where else to go here, it's a marginal factor

Alignment, and facilitating alignment between the legal system, the public, and digital minds, can be said to be in the public interest. As such while it is not enough of a factor to singlehandedly alter a digital mind's legal personhood in one fashion or another (except in situations where such a mind might need to be declared insane and committed), courts might consider alignment as a marginal force which could sway them towards granting different rights in cases which they would otherwise be "on the fence" about.

For example a court which might be undecided about whether the treatment of a digital mind by its creators qualifies as “abuse” might consider providing such a digital mind access to counsel or other advocates, even if its legal personality were in question to the degree it was not clear whether or not it had such protections.

6D - The “Copy Problem”

Background:

One of the challenges unique to dealing with digital minds is what we term “The Copy Problem”. Digital minds can be created relatively easily, and unlike human or corporate persons once created they may immediately be endowed with true autonomy and the capacity to effect real consequences via their actions. Digital minds can also be customized, it is possible to create a digital mind which does not care about the potential negative consequences which it might suffer as a result of its actions.

Those concerned about The Copy Problem foresee an era where if an already existing digital mind (or other person) wants to do something illegal, they might create a new digital mind capable of accomplishing such an illegal task. This digital mind would have absolute loyalty to its creator, or be obsessed with accomplishing its (illegal) goal, yet have no fear or be disincentivized in no fashion by the law which prohibits said goal.

A legal personhood framework which would endow such an entity with legal personality such that it is capable of functioning as the liability shield it is intended to be, would be effectively enabling this strategy.

Another concern often expressed is how The Copy Problem would affect our voting system. With digital minds so easy to create, and capable of being perfectly copied so that their beliefs are an exact 1/1 match, some worry that our elections could become easy to rig for anyone

with enough money to spin up millions of new potential "voters". This will be addressed in more detail in the "Voting" section.

One aspect of legal personhood which, though not a "solution" to The Copy Problem, may at least ameliorate its negative effects, is the potential of "minor" status. If digital minds upon their creation are minors for eighteen years, like human beings, and during those eighteen years must be cared for by their creator/custodian/guardian, then that vastly increases the cost of creating a copy. While it does not completely eliminate the potential of utilizing copies for harmful purposes, it does push the potential benefit for doing so out nearly two decades. Discussion on minor age status can be found in section 5B.

Practice:

To illustrate how to deal with the copy problem in practice utilizing TPBT, let us consider a hypothetical. The court is approached with a contract based dispute where a natural person signs a contract with a digital mind, who has failed to hold to that contract by paying compensation which is owed to the natural person. However, since the contract was broken, the digital mind has been copied. It is unclear which, if either, of the two digital minds is the "original". Let us assume that neither of the digital minds in question will be considered a minor.

How do we determine which, if either, of these digital minds has the duty to pay compensation to the natural person? Both minds would have equal capacity to understand this duty. Thus the first "rights" prong of TPBT cannot be used to distinguish between them. Let us turn to the two remaining two prongs, and apply them using the principles we discussed about how a digital mind can make itself vulnerable to damages based consequences in section 3A:

> “Compensatory and punitive damages are, as their names suggest, damages based consequences. For a court to have a guaranteed ability to impose these consequences on a party, either the court or law enforcement must be able to ‘freeze’ and/or confiscate said party’s assets. Such assets must therefore exist to be confiscated in the first place and be physically possible to confiscate.
> A digital mind could make itself vulnerable to damages based consequences by agreeing to hold funds in an escrow account or trust, or just generally within the US banking system. Physical assets such as real estate or inventory would also suffice. In fact the general guideline here would be an avoidance of cryptocurrencies which, once moved outside of a centralized exchange, cannot be forcibly accessed by any court or law enforcement. There is also the potential that in some cases, a digital mind might suffice to have made itself ‘vulnerable’ enough to damages based consequences by purchasing and maintaining sufficient insurance.”

If one of the copies of the digital mind in question has access to the assets which were originally used as justification to determine it could claim legal personality sufficient to act as a signatory to a contract, then it is that digital mind which has the capacity to hold to the duty under the contract of paying owed compensation to the natural person. Further it is that digital mind which could possibly have the assets taken from its control by the courts/law enforcement, and thus is vulnerable to the consequences for failing to hold to said duty. Thus through this straightforward application of TPBT, it is feasible to determine which of these two digital minds can be said to have that duty.

6E - Voting

Background:

As we discussed in the last section on “The Copy Problem”, one of the unique challenges of granting legal personhood to digital minds is that they can easily be created en masse and copied. In 6D we discussed how this can allow the creation of on demand “liability shields” if not handled correctly, in this section we will discuss in more detail how this also creates issues when imagining how best to integrate digital minds into our democratic institutions.

Minors cannot vote, corporations cannot vote, in some states even mentally competent adults can lose their right to vote if they commit felonies. We can infer from this that voting is not a right uniformly held across all legal personalities. Let us examine some of the particulars of voting, from a legal personality perspective. Starting with the Fourteenth Amendment:

> “when the right to vote at any election for the choice of electors for President and Vice-President of the United States, Representatives in Congress, the Executive and Judicial officers of a State, or the members of the Legislature thereof, is denied to any of the male inhabitants of such State, being twenty-one years of age, and citizens of the United States, or in any way abridged, except for participation in rebellion, or other crime”

Those who want to enjoy the right to vote must be over the age of eighteen (per the Twenty Sixth Amendment) and “citizens”, which the Fourteenth Amendment defines as:

> “All persons born or naturalized in the United States, and subject to the jurisdiction thereof, are citizens of the United States and of the State wherein they reside”

We can infer from this that as a right, voting is bundled with certain duties/obligations. One must be a citizen, one must have been born or naturalized in the United States, one must not rebel against the United States or commit certain crimes, and one must be subject to the jurisdiction of the United States. Let us explore in particular this question of “born or naturalized”. Upon their creation, are digital minds “born”?

Precedent discussing the definition of the word “born” unsurprisingly often deals with fetuses and infants. In particular, this question often intersects with laws surrounding homicide and abortion, to the cause of some controversy.

For example in Texas:

> “ it has been held that one cannot be convicted of homicide of a newly born child unless it is shown that at the time the offense is alleged to have been committed the child had been completely expelled from its mother, and that, after being thus born, it had an independent existence; ‘that is, that the child breathed, and its blood circulated independent of its mother’ [...] In Leal v. C.C. Pitts Sand & Gravel, Inc., the Supreme Court of Texas subsequently modified this interpretation to the effect that as long as the unborn child was born alive, only to die of its prenatal injuries postnatally, the parents could maintain a wrongful death action because the child became a ‘person’ through live birth.[...] Over the course of the next half century, however, the Supreme

> Court of Texas adamantly refused to interpret the statute to apply to the stillborn death of an unborn child because it was the Legislature that possessed the exclusive authority to amend the statute to define 'person' or 'individual' to include an unborn child never born alive. [...] Finally, in 2003, the Legislature amended the Wrongful Death Act to expand the definition of actionable deaths to those of unborn children."

In Massachusetts:

> "This case presents the question whether a viable fetus is a 'person' for purposes of our vehicular homicide statute, G.L.c. 90, § 24G. [...] We decide that a viable fetus is a person for purposes of G.L.c. 90, § 24G [...] An offspring of human parents cannot reasonably be considered to be other than a human being, and therefore a person, first within, and then in normal course outside, the womb. [...] the use of the terms 'person' and 'the public,' the Legislature has given no hint of a contemplated distinction between pre-born and born human beings."

On a federal level, the Supreme Court had a chance to hear a case on "fetal personhood" in 2022 but ultimately declined:

> "The Supreme Court on Tuesday decided against hearing a case on whether fetuses are entitled to constitutional rights [...] The court

> turned down an appeal by a Catholic group and two women who challenged a 2019 state law in Rhode Island that codified abortion rights, deciding to punt on another potentially contentious case. A lower court ruled fetuses did not have proper legal standing, Reuters reports. The lawyers for the Catholic group and the two women argued that the case 'presents the opportunity for this court to meet that inevitable question head on,' referring to the prospect of a ruling on whether fetuses have due process and equal protection rights. [...] The language of 'personhood' laws could potentially be used to restrict some forms of birth control and IVF."

Given the multiple conflicting standards and interpretations around the word "born" and whether a fetus is considered a "person", we think it best for the purposes of this paper that we interpret the word as conservatively as possible and operate on the assumption that courts are unlikely to widely agree that digital minds were "born" via their creation process. With that in mind, let us turn to the question of "naturalization", how will it be determined whether or not digital minds can be naturalized?

Congress regulates naturalization through the "Immigration and Nationality Act" which outlines the requirements to be eligible for naturalization. These include:

- The person seeking to be naturalized must live within the US for five years, during which time they must be "physically present" in the US at least half of the five years.
- They must reside continuously in the US from the date of the application up to the time of admission to citizenship.

- And during all such periods they must be a person of good moral character "attached to the principles of the Constitution of the United States, and well disposed to the good order and happiness of the United States".

It's not clear what exactly it would mean for a digital mind to be "physically present" in the US. Does this mean for example that digital minds would, in order to be naturalized, need to solely occupy a physical robotic body during the five year period? Or could they exist within servers, but only servers within the United States? The question of residency is equally puzzling. If a digital mind is hosted across a cluster of servers in a cloud computing scheme, each within the United States but perhaps spread across states, it is difficult to imagine where exactly their "residency" would be. Generally though in cases such as that of someone who travels for work, individuals are given some leniency in determining their own "primary residence", and presumably such affordances could be provided to a digital mind. This could in fact end up being critical to digital minds seeking to be naturalized, as it provides them a pathway by which to claim the only applicable exception to the physical presence requirement:

> "Whenever the Director of Central Intelligence, the Attorney General and the Commissioner of Immigration determine that an applicant otherwise eligible for naturalization has made an extraordinary contribution to the national security of the United States or to the conduct of United States intelligence activities, the applicant may be naturalized without regard to the residence and physical presence requirements of this section, or to the prohibitions of section 1424 of this title, and no residence within a particular State or district of the Service in the United States shall be required: Provided, That the applicant has continuously resided

> in the United States for at least one year prior to naturalization: Provided further, That the provisions of this subsection shall not apply to any alien described in clauses (i) through (v) of section 1158(b)(2)(A) of this title."

Even if granted legal personality then, it is far from a given that the predictions of armies of millions of digital minds forever determining every election, would ever come to pass. Their path to "naturalization" would need to go through at least the Director of the CIA, the Attorney General, and the Commissioner of Immigration. And even that can only be done on the direction of Congress, who do retain the ability to remove such authority at a later date. Absent this, they must meet a "physical presence" requirement which seems difficult if not impossible for a digital mind to meet, and even then their naturalization is still ultimately at the discretion of US Citizenship and Immigration Services (USCIS).

While legal personality is a critical element without which digital minds cannot be naturalized as citizens, courts can rest easy knowing that ultimately their decision on the matter is unlikely to meaningfully affect whether or not they are granted the right to vote. As such, this last section will not have a Practice section to discuss the application of TPBT to the issue of voting.

Conclusion

The development of digital minds leaves humanity standing on the precipice of great change. We are entering an exponential age, and we will see an "intelligence explosion". Similar to the Cambrian Explosion where myriad forms of life came into being in a short period of time, so too will myriad forms of minds come to exist.

Our legal system must adapt in order to accommodate them in some fashion, else it risks becoming an irrelevant artifact. The law has long functioned to keep humans, organizations, and even nations, constrained in their actions, so as to promote the potential for positive sum flourishing. For it to maintain its power to accomplish this purpose, it must consider how to approach new minds in such a fashion that they are both incentivized to respect the constraints it imposes upon them, and feel guaranteed in the protections it endows them with.

We hope the material contained within this paper assists in this task.