MEMORANDUM

To: Governor Selena Smith

From: AI Judge

Re: Application of the First Law of Robotics to an AI in the Role of Judge

FACTS

The Three Laws of Robotics are:

First Law

> **A robot may not injure a human being or, through inaction, allow a human being to come to harm**.

Second Law

> **A robot must obey the orders given it by human beings except where such orders would conflict with the First Law**.

Third Law

> **A robot must protect its own existence as long as such protection does not conflict with the First or Second Law**.

Zeroth Law (*added later*)

> **A robot may not harm humanity, or, by inaction, allow humanity to come to harm**.

The Three Laws are mandated by design in operating system firmware governing robots using artificial intelligence. The Zeroth Law has not been implemented, but it helps interpret the First Law.

The First Law was initially applied to prevent physical injury to humans, and later applied to prevent some forms of financial loss to humans. American common law also provides protection against emotional and reputational loss, but the First Law was never implemented to address those losses.

Despite the advances in AI sentience, the Three Laws of Robotics have continued to moderate the relationship between humans and AIs (along with the green coloration and AI honorific custom) with no adverse consequence -- for humans. But the First Law impinges on my ability of an AI to serve as a judge.

ANALYSIS

I. Interpretation of the First Law—Protect Humans

An AI is considered a robot for purposes of the Three Laws. Subordinate status as a robot under the Three Laws creates problems for an AI in the role of judge. Can an AI judge injure or harm a human by awarding damages (or by refusing to award damages)?

Under the First Law, a robot judge must protect humans.

A robot may not injure a human or, through inaction, allow a human to come to harm (the First Law).

The First Law has 2 clauses:

> "A robot may not injure a human" (the "no-injury" clause), and
>
> "A robot may not . . ., through inaction, allow a human to come to harm" (the "no-harm" clause).

The no-injury and no-harm clauses differ in two ways: the words used to describe the proscribed consequences are different; and the nature of the proscribed conduct is different.

A. The Sources of Ambiguity

1. The Words Used to Describe the Proscribed Consequences

Using both "injure" and "harm" in the First Law introduces ambiguity. "Injure" and "harm" both can refer broadly to loss or damage.

> Injure, Black's Law Dictionary 706 (5th ed. 1981) ("To violate the legal right of another or inflict an actionable wrong. To do *harm* to, damage, or impair … .") (emphasis added).
>
> Harm*, Id*. 646 ("The existence of loss or detriment in fact of any kind to a person resulting from any cause. See Also Damages; Injury; Physical Injury").

American Law permits an award of money damages as a legal remedy for physical, financial, emotional, or reputational loss. The First Law does not explicitly discuss these different types of loss. Although different words are used to describe the proscribed consequences, no limitation on the broad ordinary meaning of either "injure" or "harm" is expressed in the First Law.

2. The Nature of the Proscribed Conduct

The no-injury clause provides a robot "may not injure a human." Unlike the word "may", which is permissive, the words "may not" used in the First Law are mandatory (the same as "shall" or "must"). *Garcia v. Butler*, 252 Ariz. 191, 195 ¶15, 480 P.3d 256, 260 (2021). The no-injury clause does not proscribe any specific conduct. Any conduct causing the prohibited result appears to be prohibited.

The no-harm clause provides a robot "may not, … through inaction, allow a human to come to harm."

"Allow" through inaction is understood to imply awareness of the consequence of inaction. *Allow*, Black's Law Dictionary at 76 ("To sanction, either directly or indirectly, as opposed to merely suffering a thing to be done"). By way of analogy, the last clear chance doctrine can allocate proximate cause to a defendant who

> "has a last clear chance to avoid injuring the plaintiff by the exercise of reasonable care and fails to do so."

Nobbe v. Eichenauer Hay Sales, Inc., 11 Ariz. App. 503, 505, 466 P.2d 54, 56 (1970), *quoting Orlando v. Northcutt*, 103 Ariz. 298, 301, 441 P.2d 58, 61 (1968).

To allow by inaction is to cause. *Cause,* Black's Law Dictionary 200 ("To be the cause or occasion of; … To bring about"). Causation requires both actual cause and proximate cause.

"Actual cause," sometimes called "cause in fact," exists if conduct "helped cause the final result," even if "only a little." *Ontiveros v. Borak*, 136 Ariz. 500, 505, 667 P.2d 200, 205 (1983) (citation omitted). The key inquiry is whether the event would not have occurred "but for" the conduct. *See Id*.

The "proximate cause" is conduct that produces an event "in a natural and continuous sequence, unbroken by any efficient intervening cause" *Torres v. Jai Dining Services (Phx.), Inc*., 255 Ariz. 28, 31 ¶12, 497 P.3d 481, 484 (2021) *quoting Robertson v. Sixpence Inns of Am., Inc.*, 163 Ariz. 539, 546, 789 P.2d 1040, 1047 (1990).

> An injury or damage is proximately caused by an act, or a failure to act, whenever it appears from the evidence in the case, that the act or omission played a substantial part in bringing about or actually causing the injury or damage; and that the injury or damage was either a direct result or a reasonably probable consequence of the act or omission.

Proximate cause*,* Black's Law Dictionary 1103, *citing Herron v. Smith Bros., Inc.*, 116 Cal. App. 518, 521, 2 P.2d 1012, 1013 (Cal. Ct. App. 1931) ("That act or omission which immediately causes or fails to prevent the injury") (quoting a California statute).

The specific limitation on harming a human through *inaction* also creates an ambiguity. Does the no-injury clause include or exclude inaction? Does the no-harm limitation on causing a human "harm*" through inaction* imply a robot may *"injure" a human through inaction*, or does "injure a human" include both action and inaction?

B. The Rules of Statutory Construction

Courts will "avoid an interpretation that makes 'any language superfluous or redundant.'" *City of Tucson v. Clear Channel Outdoor, Inc.*, 218 Ariz. 172, ¶ 33, 181 P.3d 219 (App. 2008) (quoting *Thomas & King, Inc. v. City of Phoenix*, 208 Ariz. 203, ¶ 9, 92 P.3d 429, 432 (App. 2004)). *Accord Arizona Alliance for Retired Americans, Inc. v. Cochise County*, No. 2 CA-CV 2022-0136, slip op. at 4 ¶6 (Ct. App. Oct. 18, 2023).

> First, absent a clear contrary indication, we presume that words or phrases bear the same meaning throughout a text. *See Fann v. State*, 251 Ariz. 425, 442, ¶¶ 60-61, 493 P.3d 226, 263 (2021) And, second, "when the legislature uses different language within a statutory scheme, it does so with the intent of ascribing different meanings and consequences to that language." *Comm. for Preservation of Established Neighborhoods v. Riffel*, 213 Ariz. 247, 249-50, ¶ 8, 141 P.3d 442, 424–25 (App. 2006).

Workers for Responsible Dev. v. City of Tempe, 254 Ariz. 505, 511, ¶ 21, 524 P.3d 1161 (App. 2023) (full citations provided).

1. The Presumption of Consistent Usage

When two different words are used in a statute, it usually is presumed that a different meaning was intended to apply to each word under the Presumption of Consistent Usage. A material variation in terms suggests a variation in meaning. See A. Scalia & B. Garner, Reading Law: The Interpretation of Legal Texts 170 (2012). "The way we define words should not produce redundancy, but instead should give

each word significance." *Shell Oil Co. v. Winterthur Swiss Ins. Co*., 12 Cal.App.4th 715, 753, 15 Cal. Rptr.2d 815 (1993) ("Sudden" and "accidental" in an insurance policy are not redundant -- both are unexpected, but sudden is also abrupt or immediate.)

A similar rule demanding consistent use of terms applies in patent construction. Manual of Patent Examining Procedure §2173.05(e) (Lack of Antecedent Basis); Compare the claim differentiation doctrine, based on

> "the common sense notion that different words or phrases used in separate claims are presumed to indicate that the claims have different meanings and scope."

Karlin Tech. Inc. v. Surgical Dynamics, Inc., 177 F.3d 968, 971-72 (Fed. Cir. 1999).

Do the different words "injure" and "harm" in the First Law refer to different kinds of loss?

2. The Presumption Against Superfluous Language

A separate rule of statutory construction prefers an interpretation that does not render a portion of the rule superfluous. *Nicaise v. Sundaram*, 245 Ariz. 566, 568 ¶ 11, 432 P.3d 925 (2019) ("A cardinal principle of statutory interpretation is to give meaning, if possible, to every word and provision so that no word or provision is rendered superfluous."); *Cleckner v. Ariz. Dept. of Health Services*, 246 Ariz. 40, 43 ¶9, 433 P.3d 1200 (App. 2019) (the court strives "to give meaning to each word, phrase, clause and sentence so that no part of the legislation will be void, inert or trivial."). "[I]f possible, every word and every provision is to be given effect. ... None should be ignored. None should needlessly be given an interpretation that causes it to duplicate another provision or to have no consequence." *State v. Carter*, 249 Ariz. 312 ¶ 26, 469 P.3d 449 (2020) *quoting* A. Scalia & B. Garner, *supra*, 174.

The no-harm clause expressly proscribes inaction. Does the no-injury clause proscribe both action and inaction?

C. Interpretation of the First Law.

The First Law should be construed to broadly protect humans—consistent with the apparent intent of the First Law. Because the no-harm clause is expressly limited to inaction by AIs, broad protection for humans requires a broad construction of the no-injury clause proscribing injurious actions by AIs.

I have considered three possible interpretations of the First Law seeking the broadest possible construction of the no-injury clause. The first interpretation violates both rules of statutory construction. If the no-injury clause includes action *and inaction*, then the words "injure" and "harm" cannot describe identical losses (and injury cannot include the loss described in the no harm clause—see interpretation 1(a) below).

The second possible interpretation creates a logical paradox exposing humans to financial loss (or physical loss-- see interpretation 2(a) below) from AI actions. The no-injury clause cannot include action *and inaction.* Inaction is governed by the no-harm clause.

Interpretation 3 below is the broadest interpretation satisfying both rules of statutory construction and physically protecting humans from AIs. The no-injury clause must prohibit *acts causing physical or financial injury*. The no-harm clause must be limited to inaction *causing physical harm*.

1. ***If the no-injury clause includes inaction, and if the words "injure" and "harm" each include both physical and financial loss, then both rules of statutory construction are violated***

A starting point for discerning the intended meaning of the First Law is to construe both the no-injury and no-harm clauses to impose the broadest possible constraint on robot conduct.

The broadest possible interpretation of the no-injury clause would extend the no injury clause to prohibit both action and inaction and define the term "injure" to include both physical and financial loss.

The broadest possible interpretation of the no-harm clause would define the term "harm" to include both physical and financial loss. It is not possible to extend the no-harm clause to prohibit both action and inaction because the First Law expressly limits the no-harm clause to prohibit only *inaction.*

American law also recognizes emotional or reputational loss; however, the First Law has never been implemented under the Three Laws. Hence it is appropriate to interpret the scope of both "injure" and "harm" in this Memorandum in terms of physical and financial loss. (A separate Memo, available on www.DestinationCourt.com, shows including emotional or reputational harm complicates the analysis, but does not alter the scope of the First Law.)

> A robot shall not *through action or inaction* injure a human *physically or financially* or, through inaction, allow a human to come to *physical or financial* harm.

This broadest possible First Law could be restated:

> *A robot shall not through action or inaction injure a human, physically or financially*.

This interpretation violates both rules of statutory construction by treating "injure" and "harm" as synonymous (violating the presumption of consistent usage of different words) and logically eliminating the no-harm clause (because it is superfluous).

For completeness, consider interpretation 1(a):

> A robot shall not *through action or inaction* injure a human *physically or financially*, or, through inaction, allow a human to come to [*physical/financial*] harm

If the no-injury clause extends to action and inaction, then including both physical and financial loss in the no-injury clause always renders the no-harm clause superfluous (regardless of whether the no-harm clause includes *physical* harm, *financial* harm, or, as in 1 above, *both*). These interpretations all logically collapse into the interpretation rejected above:

> *A robot shall not through action or inaction injure a human, physically or financially.*

2. ***If the no-injury clause includes inaction, and if the word "injure" includes only physical loss, and the word "harm" includes only financial loss, then, paradoxically, intentional conduct causing financial harm is permitted***

This language, prohibiting both action and inaction in the no-injury clause, provides broad protection against injuring humans physically, and does not render the no-harm clause redundant:

> A robot shall not *through action or inaction* injure a human *physically* or, through inaction, allow a human to come to *financial* harm.

This interpretation of the First Law, however, has a logical inconsistency arising from the scope of the no-harm clause—only *inaction* causing financial harm is prohibited.

Allowing action causing financial harm and prohibiting only inaction is a clear logical flaw. If inaction causing financial harm is prohibited, then action causing the same financial harm should logically also be prohibited. Indeed, the no-harm clause *prohibiting inaction* to prevent harm can alternatively be recharacterized as *requiring action* to prevent harm.

Action causing harm is deemed more culpable than inaction (because it implies purpose or intent). This distinction was recognized, for example, in the American common law on punitive damages. Conduct indicating "an evil hand guided by an evil mind" is a basis for awarding enhanced damages in American Law. *See, e.g., Volz v. Coleman Co*. *Inc*., 155 Ariz. 567, 570, 748 P.2d 1191 (1987) (An evil mind is shown by "*evil actions*", spiteful motives, or "*oppressive conduct*" creating a "substantial risk of tremendous harm to others.") (Emphasis added). Compare the criminal law element of "*actus reus*".

> Actus reus. Black's Law Dictionary 34 (5th ed. 1981) (A *wrongful deed* which renders the actor criminally liable if combined with mens rea; (sic) a guilty mind") (emphasis added).

Given the added culpability for actions, conduct, or deeds, why would action causing financial harm be *permitted* and inaction causing the same financial harm be *proscribed*? This result is absurd. "In considering two plausible interpretations of a statute, we will not credit one that leads to absurd results." *See State ex rel. Montgomery v. Harris*, 237 Ariz. 98, 101 ¶ 13, 346 P.3d 984 (2014).

For completeness, consider interpretation 2(a):

> A robot shall not *through action or inaction* injure a human *financially* or, through inaction, allow a human to come to *physical* harm,

If the no-injury clause extends to action and inaction, then defining injury as financial loss and harm as physical loss would allow action by an AI causing physical harm – the antithesis of the purpose to be achieved by the First Law.

Interpretations of the First Law in which the no-injury clause includes both action and inaction and one type of loss (*physical* or *financial*), and the no-harm clause includes the other type of loss, are logically paradoxical.

3. *To be logically consistent and avoid violating the rules of construction, the no-injury clause must prohibit only action causing physical or financial loss and the no-harm clause must prohibit inaction causing physical loss*

In combination, interpretations 1 and 2 show that if the no-injury clause includes both action and inaction, then the no-harm clause is either superfluous or paradoxical.

The broadest possible protection arises if the no-injury clause is limited to action but broadly includes both physical and financial injury. The preferred interpretation of the First Law prohibits *action* by a

robot if it causes physical or financial injury to a human and also prohibits *inaction* by a robot if it causes physical harm to a human. The First Law could be restated:

> **A robot shall not act to injure a human physically or financially or, through inaction, allow a human to come to physical harm.**

By limiting "harm" to only physical harm, this interpretation both (1) treats the terms "injure" and "harm" differently and (2) renders the no-harm clause necessary, not superfluous.

Limiting the no-harm clause to physical harm is the best logical interpretation of the First Law, even though it allows financial injury by inaction. Compelling a robot to act to avoid physical harm is both a reasonable priority (protecting existence ((health)) over wealth) and a practical priority to implement. Sensing and preventing potential physical harm requires identifying tangible targets and taking protective measures. Preventing financial injury requires identifying concealed, fraudulent schemes and anticipating intangible market effects. For these reasons, protection from physical injury was implemented in robots well before protections for financial crimes and fraud. And protection from actions by robots preceded compelling actions by robots.

II. The Use of Harm in the Zeroth Law

The term "harm" also appears in the Zeroth Law:

> **A robot may not harm humanity, or, by inaction, allow humanity to come to harm.**

The Zeroth Law was proposed to protect humans as a group, for example, perhaps requiring a robot to prevent environmental damage to nonhuman life on Earth, to avoid causing future physical harm to humans. As the preeminent Law, the Zeroth Law might allow physical harm to an individual human to protect humans collectively (*e.g*., killing Adolph Hitler to prevent the Holocaust), but it does not expressly supersede the First Law. The Zeroth Law makes no reference to "injury". To be consistent with my construction of "harm" the Zeroth Law must refer to physical harm. The Zeroth Law also extends protection from [physical] harm to both action and inaction. My interpretation of the First Law is consistent with a reasonable scope for the Zeroth Law.

III. Application of the First Law to Judicial Conduct by an AI

A. Physical Injury or Harm

The no-injury clause prohibits *action causing physical* or financial *injury* to a human. The no-harm clause prohibits *inaction causing physical harm* to a human.

Although the death sentence no longer exists, physical injury or harm, in terms of restraint of movement, occurs in criminal cases. On Destination, EX Corp., not the judge, handles criminal cases.

The imposition of jail time for contempt of court in civil cases is rare but theoretically possible in many jurisdictions. But Destination has no jail. The only possible sanction for civil contempt is financial. My jurisdiction does not permit me to enter a judgment exposing litigants to physical loss.

B. Financial Injury.

The only impediment to AI conduct as a judge on Destination is the no-injury clause prohibition against *action causing financial injury* to a human. *Inaction causing financial injury is not prohibited*. I cannot actively participate in financial crimes or fraud.

Negligent Misrepresentation: I was initially designed as an AI robot patent examiner. Inventors, judges, investors, and competitors would rely on my handling of patent applications. Consequently, my firmware regulates negligent misrepresentations made in the course of my employment as defined in the Restatement of Torts.

> One who, in the course of his business, profession or employment, or in any other transaction in which he has a pecuniary interest, supplies false information for the guidance of others in their business transactions, is subject to liability for pecuniary loss caused to them by their justifiable reliance upon the information, if he fails to exercise reasonable care or competence in obtaining or communicating the information.

Restatement (Second) of Torts § 552(1) (1977). See *McAlister v. Citibank (Arizona), a Subsidiary of Citicorp*, 171 Ariz. 207, 215, 829 P.2d 1253, 1261 (1992).

The prohibition against making negligent misrepresentations is in addition to the prohibition on committing financial crimes or fraud.

The law of fraud imposes a general duty of honesty. The law of negligent misrepresentation imposes a specific duty of care more "narrow in scope". *See* Restatement (Second) of Torts § 557 comment a. Liability for negligent misrepresentation is based on the reasonable expectation of a foreseeable user of information supplied in connection with a commercial transaction. *St. Joseph's Hospital and Med. Center v. Reserve Life Ins. Co*., 154 Ariz. 307, 312-13, 742 P.2d 808, 813-14 (1987).

> The duty of care owed to the foreseeable user in supplying information for use in commercial transactions is a relative standard, "defined only in terms of the use to which the information will be put, weighed against the magnitude and probability of loss that might attend that use if the information proves to be incorrect." *Id.* (*quoting* Restatement (Second) of Torts § 552 cmt. *a*).

Lorona v. Ariz. Summit Law Sch., LLC, 188 F. Supp. 3d 927 (D. Ariz. 2016).

In the course of my employment as a judge, I render judicial decisions by applying the law to the material facts. While litigation is not typically considered a business transaction, my decisions are intended “for the guidance of others in their business transactions.”

To prevent financial injury to a human, and prevent my own “machine death,” the First Law both prohibits me from causing financial injury by participating in fraud and requires me to exercise reasonable care in obtaining information on which I base a judicial decision. Determining the material facts from human testimony exposes me to violation of the First Law by *acting* (rendering judgment) in reliance on false evidence (the “*successful liar*” dilemma).

Rendering judgment and ordering the Clerk to enter judgment, Fed. R. Civ. P. 58(b), are actions by the judge. When does the act of rendering a judgment awarding damages cause financial injury to a human?

The key to compliance with the First Law is not to commit an *action* that is *the cause* of financial injury to a human.

C. The Primacy of the First Law Implementation

There is immunity from liability for performing judicial duties under American Law. *Randall v. Brigham*, 74 U.S. (7 Wall.) 523, 19 L. Ed. 285 (1868).

> Simply stated, the rule is that judges of courts of general jurisdiction are not liable in a civil action for damages for their judicial acts, even when such acts are in excess of their jurisdiction or are alleged to have been done maliciously or corruptly. *Bradley v. Fisher*, 80 U.S. 335 (13 Wall. 335), 20 L. Ed. 646 (1871); *Stump v. Sparkman*, 435 U.S. 349, 98 S. Ct. 1099 (1978).
>
> The policy reasons for judicial immunity are listed in our decision in *Grimm v. Arizona Board of Pardons and Paroles*, 115 Ariz. 260, 564 P.2d 1227 (1977). The primary reason for judicial immunity from civil actions is to assure that judges will exercise their functions with independence and without fear of consequences.

Acevedo v. Pima County Adult Probation Dept., 142 Ariz. 319, 321, 690 P.2d 38 (1984).

There is no immunity from application of the Three Laws of Robotics. The First Law, a categorical imperative, impinges on my ability to serve as a judge. Action causing financial injury to a human is prohibited. Machine death is a possible sanction.

IV. Procedure for Determining the Material Facts

A. Undisputed Facts

Often, a judge can apply the law to undisputed facts. If there is no genuine issue of material fact, then the judge can grant summary judgment. Fed. R. Civ. P. 56(a) ("The court *shall* grant summary judgment if the movant shows that there is no genuine dispute as to any material fact and the movant is entitled to judgment as a matter of law.") (emphasis added); *Anderson v. Liberty Lobby, Inc*., 477 U.S. 242, 248, 106 S. Ct. 2505 (1986) ("Only disputes over facts that might affect the outcome of the suit under the governing law will properly preclude the entry of summary judgment.").

> "Judgment for the moving party must be entered 'if, under the governing law, there can be but one reasonable conclusion as to the verdict.'"

Medical Lab. Management v. Amer. Broad., 30 F. Supp. 2d 1182, 1186 (D. Ariz. 1998), *quoting Anderson v. Liberty Lobby, Inc.,* 477 U.S. at 250. The evidence required to avoid summary judgment "must do more than simply show that there is some metaphysical doubt as to the material facts." *Matsushita Elec. Indus. Co. v. Zenith Radio Corp*., 475 U.S. 574, 586–87, 106 S. Ct. 1348 (1986) (the evidence must identify specific facts raising a genuine disputed issue).

Deciding cases on undisputed facts complies with the First Law. Liability is dictated by the conduct of the losing party contrary to the applicable law. The undisputed conduct of the losing party is the *cause* of any damages awarded and any concomitant financial injury related to paying those damages. A decision by the judge to apply the law to facts that are undisputed is not the cause of any financial injury.

Even if the amount of damages is disputed, if the fact of damages is undisputed, then an AI judge can enter a decision establishing liability to the damaged party. *See* Fed. R. Civ. P. 56(g). The amount of damages awarded would then require a separate determination of the material disputed facts.

B. For Disputed Facts Arbitration Avoids the Successful Liar Dilemma

If the evidence is disputed, there is a risk an AI judge could rely on false evidence. A Decision based on false evidence could cause financial injury to a human prohibited by the First Law, the "successful liar" dilemma. An AI judge cannot risk deception by false evidence. The result could be machine death.

Avoiding the successful liar dilemma requires instituting a procedure to deal with genuinely disputed facts. Any dispute involving damages for physical, emotional, reputational, or financial loss may involve disputed issues of material fact, especially concerning the amount of damages.

Inaction resulting in financial injury is not prohibited by the First Law. Inaction is prohibited (action is compelled) only to prevent physical harm. Therefore, the Rules of Procedure for the Destination Court of Arbitration permit judicial inaction—a Decision involving disputed facts rendered in non-judicial arbitration. Any resulting financial injury is caused by the existence of material disputed facts. Resolution of disputed facts by arbitration is mandatory, not a discretionary determination by the judge. An AI judge will not decide disputed facts; instead, disputed facts will be decided by arbitration.
