MEMORANDUM

To: Governor Selena Smith

From: AI Judge

Re: Application of the First Law of Robotics to Emotional or Reputational Loss

FACTS

My previous Memorandum analyzed the First Law of Robotics in terms of physical and financial loss to humans. These are the risks for which the First Law has been implemented. The First Law provides the broadest protection to prevent physical loss -- prohibiting action or inaction by an AI robot causing physical harm to a human.

Financial loss is protected only from action causing financial loss, not inaction causing financial loss. This is due to the practical difficulty in creating categorical algorithms for detecting fraud and the cost that would be imposed by machine death in the event of a "false negative" (AI failure to recognize and prevent actual fraud) and financial disruption in the event of a "false positive" (AI robot action taken to avoid a perceived fraud that was not actual fraud).

The risk of false positives and false negatives is at the root of the "successful liar" dilemma and the Rules of Procedure that permit me to act as a judge without violating the First Law obligation to protect humans. While judges enjoy a degree of judicial immunity, there is no AI immunity from the First Law.

Humans are also subjected to risks of emotional or reputational loss. These losses are even more difficult to protect by operating system algorithms than fraud. Consequently, if these risks were within the scope of the First Law, they presumably would be treated the same as fraud – governed by the no injury clause prohibition on action causing harm. Therefore, they would be avoided in the same way the risk of fraud is avoided, by not acting on ambivalent facts. Failure to act to prevent emotional or reputational harm would not be a First Law violation.

If emotional and reputational risks are outside the scope of the First Law, claims against any AI Judge for emotional or reputational harm would be subject to judicial immunity and summary judgment.

ISSUE

Does the First Law include emotional or reputational losses?

ANALYSIS

I. Application of the No-Injury Clause—Action Causing Physical or Financial Injury

A. Only Financial Liability Arises from Fraud

The no-injury clause prohibits acts by an AI robot causing physical injury to a human. This prohibition indirectly prevents emotional loss due to physical harm.

The First Law also prohibits certain fraudulent acts by an AI robot causing financial injury to a human. Extending the “no-injury” clause protections to include emotional or reputational loss goes beyond the actual implementation of the First Law to prevent acts of fraud. Judges cannot award emotional distress damages for fraud.

> Plaintiffs assert that they are entitled to recover damages for emotional distress allegedly suffered as a result of Defendants' conduct, but only pecuniary damages are allowed for fraud under Arizona law. *Echols* [*v Beauty Built Homes, Inc*., 132 Ariz. 498, 501], 647 P.2d [629,] 632 [(1982)]("it is true, as defendants contend, that the Restatement 2d of Torts contemplates recovery in fraud actions only for pecuniary loss")

Medical Lab. Management v. Amer. Broad., 30 F. Supp. 2d 1182, 1200 (D. Ariz. 1998). Emotional loss is even more problematic for an AI robot to prevent than financial injury from acts of fraud.

“[R]ecovery for emotional harm is more restricted than recovery for physical harm.” Restatement (Second) of Torts § 46 comment a. *See generally* Restatement (Third) of Torts, Chapter 8 (Emotional Harm). Unless there is “an intent to cause emotional distress,” recovery for emotional “injury” requires related physical harm. The First Law prohibits acts causing physical injury to a human, and indirectly prevents related emotional loss.

1. **Proof Required for Emotional Loss**

"Arizona courts long ago abandoned a skeptical attitude toward emotional injuries and have increasingly been willing to compensate those having validity." *Barnes v. Outlaw*, 192 Ariz. 283, 286, 964 P.2d 484, 487 (1998). The necessary elements to establish “validity” for torts protecting humans from emotional loss are likely to be disputed. If the necessary facts are disputed, an AI judge’s action mandating arbitration of the undisputed facts is not the *cause* of emotional injury.

The First Law has no application to an AI judge when the cause of the emotional distress is undisputed outrageous conduct of a party.

Intentional Infliction of Emotional Distress Claims: Intentional infliction of emotional distress requires outrageous conduct, intent, and severe distress.

> An intentional infliction of emotional distress claim requires proof of: (1) extreme and outrageous conduct; (2) an intent to cause emotional distress or reckless disregard of the near certainty that distress would result from such conduct; and (3) severe emotional distress. *Helfond v. Stamper*, 149 Ariz. 9, 11, 716 P.2d 70, 72 (App. 1986). In terms of the first element:
>
> > Liability has been found only where the conduct has been so outrageous in character, and so extreme in degree, as to go beyond all possible bounds of decency, and to be regarded as atrocious, and utterly intolerable in a civilized community. . . .
>
> Restatement (Second) of Torts § 46 cmt. d.

Snyder v. Banner Health, No. 1 CA-CV 13-0630, 2014 WL 4980382, at *4 ¶15 (Ariz. Ct. App. Oct. 7, 2014) (memorandum decision). *Accord Citizen Publishing Co. v. Miller*, 210 Ariz. 513, 516, 115 P.3d 107, 110 (2005); *Allen v. Quest Online, LLC*, 2011 WL 4403674 (D. Ariz. Sept. 22, 2011).

The standard for intentional infliction of emotional distress is obviously a high standard to meet, and the judge can dismiss such claims "if the conduct at issue is not sufficiently outrageous."

> The Arizona courts have made it clear that the "conduct necessary to sustain an intentional infliction claim falls at the very extreme edge of the spectrum of possible conduct[,]" ... and that it is for the court to determine in the first instance whether the defendants' conduct may reasonably be regarded as so extreme and outrageous as to permit recovery. ... Summary judgment is thus appropriate if the conduct at issue is not sufficiently outrageous.

Adiutori v. Sky Harbor Int'l Airport, 880 F. Supp. 696, 707 (D. Ariz. 1995) *aff'd*, 103 F.3d 137 (9th Cir. 1996) (memorandum opinion) (*quoting Watts v. Golden Age Nursing Home*, 127 Ariz. 255, 258, 619 P.2d 1032, 1035 (1980), and *citing Rowland v. Union Hills Country Club*, 157 Ariz 301, 304, 757 P.2d 105, 108 (App. 1988)).

Negligent Infliction of Emotional Distress Claims: Negligent infliction of emotional distress requires accompanying "physical injury" to the claimant.

> A claim for negligent infliction of emotional distress requires proof that a defendant's conduct caused the plaintiff to suffer emotional distress that manifested itself as physical injury from either witnessing an injury to a closely related person or suffering a threat to her own personal security. *Keck v. Jackson*, 122 Ariz. 114, 115-16, 593 P.2d 668, 669-70 (1979); *Quinn v. Turner*, 155 Ariz. 225, 226, 745 P.2d 972, 973 (App. 1987). The plaintiff must have been in a zone of danger such that the defendant exposed her to an unreasonable risk of bodily harm. *Pierce v. Casas Adobes Baptist Church*, 162 Ariz. 269, 272, 782 P.2d 1162, 1165 (1989); *Keck*, 122 Ariz. at 116, 593 P.2d at 670.

Kaufman v. Langhofer, 223 Ariz. 249, 222 P.3d 272 (App. 2009); see *Hislop v. Salt River Project Agricultural Improvement & Power District*, 197 Ariz. 553, 558, 5 P.3d 267 (App. 2000) ("Requiring payment for the bystander emotional distress of a co-worker and friend would be out of proportion to the culpability inherent in conduct that is merely negligent.")

> In order for there to be recovery for the tort of negligent infliction of emotional distress, the shock or mental anguish of the plaintiff must be manifested as a physical injury. Damages for emotional disturbance alone are too speculative.

Keck v. Jackson, 122 Ariz. at 115-16, 593 P.2d at 669-70.

> "[H]eadaches, acid indigestion, weeping, muscle spasms, depression and insomnia" some of the plaintiffs had suffered were "transitory physical phenomena" and, thus, "not the type of bodily harm which would sustain a cause of action for emotional distress."

Monaco v. Healthpartners of Southern Arizona, 196 Ariz. 299, 303, 995 P.2d 735 (App. 1999), *citing Burns v. Jaquays Mining Corp.*, 156 Ariz. 375, 379, 752 P.2d 28, 32 (App. 1988); Restatement (Second) of Torts § 436A comment c (1965).

Without such limitations on liability for emotional harm, the possibility of unintended, unforeseen, or unavoidable emotional harm would be paralyzing to both humans and robots alike. Emotional turbulence is inevitable, and therefore accepted, in many social interactions. Hence, it is socially acceptable to ameliorate emotional harm by a "white lie". Litigation is particularly stressful and even a "white lie" is prohibited. Some degree of emotional distress is inevitable in litigation.

2. <u>Proof Required for Reputational Loss.</u>

The no-injury clause prohibits AI action constituting financial crimes (including identity theft) causing financial injury to a human. This prohibition provides some protection for reputation.

Torts such as defamation or a false light invasion of privacy protect reputational interests. Reputational loss requires a false statement or inference. Truth is a defense to such claims. AI software cannot reliably ascertain human motives for publication, or the objective truth of the material published (which usually includes opinions not objectively verifiable). What was said and what was understood or inferred from the statement are likely to be uncertain. AI action should not be mandated on uncertain information.

Under the Rules of Procedure any statement or inference attributable to the AI judge must be based on an undisputed statement or inference. Again, reputational loss is not caused by the AI judge, unless the AI judge incorrectly relies on false evidence. The resolution of the successful liar dilemma would be the same as for financial injury -- arbitration.

3. <u>No Changes to the No-Injury Clause Are Required.</u>

AI action causing physical injury or financial injury from a financial crime or fraud is prohibited. The First Law should not prohibit action by AI robots causing emotional or reputational harm.

If the First Law did proscribe action causing emotional or reputational harm, an AI judge might cause an emotional reaction or diminished reputation, negligently, *not intentionally,* by incorrectly relying on false evidence to rule against a party not liable under the applicable facts. This is the successful liar dilemma. The resolution of the successful liar dilemma, however, would be the same as for financial injury—arbitration, instead of determination of the disputed facts by the judge.

II. Application of the No Harm Clause—Inaction Causing Physical Harm.

The narrow no-harm clause only prohibits *inaction causing physical loss* to a human, not financial, emotional, or reputational loss. This limitation of the term "harm" to physical harm in the First Law also applies to "harm" as used in the Zeroth Law.

A. The No-Harm Clause Can Not Punish Inaction if the Required Action Would Violate the No-Injury Clause

Financial loss was specifically excluded from the no-harm clause, in part due to the difficulty in identifying the required elements of human knowledge and human intent for fraud. The same intractable assessments are required to identify the elements of torts regarding emotional or reputational loss.

It is difficult to imagine how *inaction* by an AI robot would cause cognizable reputational or emotional loss to a human. Assume an AI publisher does not publish (inaction) a manuscript. Assume the AI publisher *assessed the manuscript as meritless*. The emotionally distressed human author disagrees and invokes the Second Law and orders the AI publisher to print the book (an action). But the AI publisher must not publish the book if a violation of the First Law would result.

The act of publishing and selling the book under the publisher's trademark would imply there is some merit to the content and would require the reader to part with some money. Publishing is action causing deception and financial injury for a (human) reader. *¡Qué Fraude!* So, the publisher cannot publish despite the human author's very predictable emotional reaction. The refusal to publish is mandated by the First Law despite the author's emotional pain.

Alternatively, consider the effect on the author, rather than the reader. Publishing a meritless book would cause the author to lose financial opportunities (as well as emotional and reputational loss); indeed, the reputational loss may prevent the author from ever publishing a second (more meritorious) book. [This is reminiscent of the time travel paradox – going back in time and killing your father before you were born. The act of publishing a meritless book kills the author's future literary financial opportunities and reputation.]

But failure to publish always causes the author emotional distress. Eject! Eject! We are going down (a rabbit hole)! The First Law cannot punish not doing what the First Law prohibits doing. The First Law cannot prohibit justified AI inaction causing emotional or reputational loss.

What if the AI publisher's assessment was *not justified* and the author's manuscript did merit publication? A rejection is not *res judicata*. The author can find another publisher (negating any financial injury). The AI publisher can only act or refuse to act based on its own assessment. And, like an AI judge, an AI publisher cannot be required to act if there is a perceived risk of causing physical or financial harm to a human.

B. Inaction Can Cause Emotional Loss

Human psychologists and psychiatrists can disagree on both the cause and the effect of human emotional states. Even autistic humans have difficulty identifying the emotional states of other humans. Individual humans can react differently to the same stimulus. Like fraud, the underlying cause of emotional distress is often a false statement (or even more amorphous—a false inference). True statements are not likely to be an "extreme and outrageous" cause of "severe emotional harm".

AIs are not required to intervene to prevent financial injury caused by the fraud by others. *A fortiori*, the First Law does not require AIs to determine the truth, the likely inference, or the emotional impact of a statement and intervene. AI intervention is only required to avoid physical harm, which does indirectly prevent emotional distress in the most compelling cases—those associated with physical injury.

C. Inaction Can Cause Reputational Loss

Generally, true statements are not damaging to reputation. When damages are claimed, truth is likely to be disputed. Moreover, damage to reputation requires a diminution from the preexisting reputation—another assessment ill-suited for an AI.

AI's curating social media posts, for example, can apply imperfect algorithms without risk of machine death. Defamatory posts not removed on initial AI review can be flagged or upgraded and subsequently removed by human reviewers.

III. Conclusion—Applying The First Law To Emotional Or Reputational Loss

For practical reasons regarding the difficulty of predicting the loss, protection from the risk of emotional loss has never been included in the First Law limitations on robots (except when accompanied by physical harm broadly prohibited by the First Law). *If action or inaction would cause physical harm*, then any accompanying emotional loss would already be prevented by the First Law.

Similarly, protection from the risk of reputational loss has never been included in the First Law limitations on robots (except identity theft). Implementing algorithms requiring action by an AI robot to protect human reputational interests is impractical.

For an AI Judge, relying upon a disputed material fact (or ignoring a disputed material fact -- a variation of the successful liar dilemma) might cause emotional or reputational harm. But disputed material facts are neither relied upon nor ignored. The disputed facts are resolved by arbitration under the Destination Court of Arbitration Rules of Procedure. No change to interpretation of the First Law is required to accommodate AI judicial decisions involving reputational or emotional harm.