בס"ד

חבורת מהות היהדות
Weekly Newsletter

Vol. 2, Issue 7

פרשת ויקהל
כ"ב אדר תשפ"ה

Feedback? contact@essenceofjudaism.com

From the Chabura
By: Adam Friedmann

The Alignment Problem: *Iyov* and *Get Meusah*

As AI systems become more prominent in our lives, an increasingly urgent question emerges: How do we ensure that AI behaves in ways that align with human values?

This question, known as the Alignment Problem, refers to the challenge of designing AI systems that act in accordance with human goals, ethical principles, and values. While AI is designed to optimize for specific objectives, there are countless ways in which these objectives can be misaligned with human intentions. The consequences of misalignment range from minor inconvenience to catastrophe. To illustrate, let's consider three types of misalignment issues that have already surfaced in real-world AI systems:

1. Insufficient data leading to model failures
AI models rely on vast amounts of data to function effectively. However, when a model lacks sufficient or representative data, it can lead to serious errors. One infamous example occurred in an early version of Google's image classification AI, which mistakenly labeled images of black people as “gorillas.” This failure was not due to intentional bias in the AI but rather a lack of diversity in the images used to train the model.

2. AI models reflect data, not reality
Another major source of misalignment arises because AI models are based on human-generated data, which may itself be biased or incomplete. An early AI language model, Word2Vec, demonstrated how language-processing systems may inadvertently encode societal biases. For example, Word2Vec, trained on a large corpus of text, learned to associate words like “man” with “programmer” and “woman” with “homemaker.” Do these word relationships reflect something fundamental about reality, or are they systemic biases that only exist in the language people produce?

3. AI optimizing for the wrong goals
The most pressing concern in AI alignment is the risk that AI systems, if given control over complex processes, may optimize for goals that are misaligned with what is best for humans. As AI systems take on greater roles in banking, corporate decision-making, and resource allocation, misalignment could lead to unintended consequences. For instance, an AI-driven financial system optimizing purely for profit might engage in high-risk strategies that destabilize markets, while an AI managing a supply

chain could make decisions that prioritize efficiency over ethical considerations.

A philosophical challenge
At the heart of the AI alignment problem lies a deeper philosophical question: Are humans even capable of determining what their true needs and values are? If AI is to be aligned with human interests, we must first understand what those interests truly entail. Yet, the Jewish tradition suggests that human beings are inherently limited in their ability to grasp their own needs.

***Iyov*: human limitations on understanding reality**
One of the most profound discussions of human limitations can be found in *Sefer Iyov*. Throughout the narrative, Iyov seeks to understand the reasons behind his suffering. One of Hashem's ultimate responses (*Iyov* 38-39) is that Iyov's understanding of the world is myopic. Human beings can't grasp the workings of the universe and therefore can't understand their true place within it. Without this understanding, can we hope to create AI systems that will optimize human flourishing?

***Get meusah*: the conflict of human drives**
A more psychological concern is codified in halachah. Generally, a husband must grant a get of his own free will. A compelled divorce (*get meusah*) is invalid. However, the Mishnah (*Gittin* 9:8) rules that in certain cases, a Beit Din can compel a husband to grant a divorce, and this is considered valid (see *Gittin* 88b and *Bava Batra* 48b-49a). The Rambam (*Hilchot Gerushin* 2:20) explains that this works because the recalcitrant husband has conflicting internal drives: on the one hand, he wants to do what is right according to halachah and to maintain his place in the Jewish community; on the other, his *yetzer hara* pushes him to resist. When the court compels him, it is merely weakening his baser drive so that his true will can emerge.

This concept reflects a fundamental truth about human psychology: people are often driven by conflicting desires, and at times, their baser instincts prevent them from recognizing their own best interests. If humans struggle to determine their true will in their own lives, how can they be trusted to define the optimal outcomes for AI alignment? These examples highlight a critical dilemma. The very concept of "AI alignment" presumes that we can discern what is truly the best for us. But our *mesorah* indicates that achieving this is out of our grasp.

This is why, according to Jewish tradition, humans require a higher-order level of guidance to navigate the complexities of life. The Torah serves as this guiding force, providing a moral and ethical framework that transcends human biases and limited perspectives (see *Chovot Halevavot*, *Gate of Serving G-d*, 2). Just as the Torah helps humans align themselves with a higher purpose despite their internal conflicts, any attempt to align AI with human values may require principles that go beyond the limitations of human perception and short-term desires.

This presents a significant challenge for AI development. If human beings cannot fully trust themselves to make optimal decisions, should AI be designed to follow their explicit instructions, or should it be guided by some higher set of principles? And if so, who defines those principles? These are not just technical questions but profound ethical and philosophical dilemmas that must be addressed as we move toward an AI-integrated future.

Mishnah: A Philosophy of Life
By: Dovid Campbell

Nedarim 3:11 — The Greatest Mitzvah of All

The Mishnah in *Nedarim* (3:11) describes the value of the mitzvah of *milah*, with various sages taking unique approaches to its significance. Their statements cast *milah* as not only a defining marker of Jewish identity, but also as a source of spiritual unity, and even a fundamental pillar of creation itself. *Tiferet Yisrael* expands on these ideas, showing how *milah* transforms both the individual and the world.

Rabbi Yishmael states, "**Great is *milah*, for thirteen covenants were established upon it.**" *Tiferet Yisrael* explains that these thirteen covenants parallel the thirteen attributes of Divine mercy and correspond numerically to *echad* (one). Through *milah*, Jews recognize each other as bound by a sacred covenant, and they remain united with God as His servants. Unlike most mitzvot, which are performed at specific times, *milah* is an enduring physical sign of this relationship, lasting until the body's end. This permanence ensures that the covenant is never severed, making *milah* not just a personal commitment but a continuous connection to Jewish history and destiny.

Rabbi Yosi teaches, "**Great is *milah*, for it overrides even the stringent Shabbat.**" Despite Shabbat's centrality to Jewish life, *milah* takes precedence, demonstrating that some mitzvot, though fundamentally holy, must yield to an even greater principle. Shabbat sanctifies time, but *milah* sanctifies the body itself, embedding holiness into the very being of a Jew.

Rabbi Yehoshua ben Karchah observes, "**Great is *milah*, for it was not suspended even for Moshe.**" *Tiferet Yisrael* explains that Moshe was punished immediately for postponing his son's circumcision, even though the Torah often allows for delays in performing mitzvot when justified. However, *milah* is different—it is a *mitzvah of the body*, defining not only what a person does but who they are. A Jew's covenant with God is not a matter of convenience or circumstance; it is essential to their very existence.

Rabbi Nechemiah states, "**Great is *milah*, for it overrides *tzara'at*.**" *Tiferet Yisrael* notes that even a delayed *milah* (past the eighth day) still takes precedence over the prohibition of cutting a *tzara'at* lesion. In a sense, *milah* is not just a commandment but a force of purification, overriding even ritual impurity. Similarly, Rabbi Yehuda HaNasi teaches, "**Great is *milah*, for even with all the mitzvot that Abraham performed, he was not considered complete until he was circumcised**" (*Bereishit* 17:1). The transformative power of *milah* to create spiritual wholeness fundamentally transcends all other commandments.

The Mishnah concludes with an even more striking statement: "**Great is *milah*, for were it not for *milah*, the Holy One, blessed be He, would not have created His world.**" The Mishnah cites *Yirmiyahu* 33:25 as its source: "Were it not for my covenant, day and night, I would not have appointed the statutes of heaven and earth." *Tiferet Yisrael* explains that the verse is emphasizing the constancy of *milah*, which uniquely remains with the body day and night. The implication is that it is specifically this quality of constancy in our Divine service that justifies creation—a constancy we uphold whenever we contemplate our enduring covenant with our Creator.

Sforno on the Parsha
By: Nochum Spiegel

A Palace in Time

The diligent performance of *mitzvot* is our sure path to spiritual perfection. What of a scenario where two commandments collide, where the fulfillment of one precludes the execution of a second? Chazal's comment to the beginning of our *parsha* addresses one of these scenarios.

Moshe's instructions regarding the building of the *Mishkan* are prefaced by "six days work shall be done, and the seventh day shall be holy for you, a Shabbat of complete rest to Hashem" (*Shemot* 35:20). *Mechilta DeRabbi Yishmael* explains that we may have thought that the construction of the *Mishkan* should continue unabated seven days a week. "Six days work shall be done" informs us otherwise. Even pursuit of the great spirituality resulting from the presence of a *Mishkan* cannot override the holiness of the Shabbat. Sforno to *Ki Tisa* (31:13-15) explains the *pesukim* there as offering multiple reasons why this is so. We will focus on the first. "But my *Shabbats* you shall keep, **for it is a sign (*ot*) between Me and you** throughout your generations **to know that I am Hashem who sanctifies you**". Observing *Shabbat* is the foundational sign of Hashem's relationship of *Shechina* and *Kedusha* with *Am Yisrael*. One cannot build a physical edifice of *Kedusha* (i.e. *Mishkan*) by shattering the bedrock upon which it stands. To build a *Mishkan* on Shabbat is to undermine and remove the spiritual receptacle we are seeking to fill (Sforno 31:13).

In *parshat Teruma* we presented Sforno's view that the *Mishkan* was a rescue plan to uplift *Bnei Yisrael* after losing their spiritual standing due to the *Chet HaEgel*. The ideal state does not limit kedusha to a specific geographical location rather "in every place where My name is proclaimed, I will come to you and bless you" (Sforno *Shemot* 20:20). With this context perhaps we can explain the Shabbat as follows. Shabbat represents the dimension of *Kedusha* in the realm of time. It's inception from the beginning of creation as, "God blessed the seventh day and made it holy" (*Bereishit* 2:3). Its reach is not limited to a specific area or landmass. No matter where on the globe they may be or whatever their spiritual state, *Am Yisrael* has the ability to assert their dedication to this sign (*ot*) between them and Hashem. It continues from generation to generation uninhibited by the surrounding environment.

The introduction of the *mitzvah* of building a *Mishkan* represents a reality of *Kedusha* directed to a specific space. *Bnei Yisrael*, post *Chet Haegel*, are now in need of this extra dimension of focused *Kedusha*. Their spiritual levels will now be dependent on interacting with physical objects and stimuli.

In a world devoid of *Mishkan* and *Mikdash*, our *kedusha* in time still remains. In spite of exile and war, Shabbat as the underpinning of spiritual life can never be wrenched from the hearts and souls of the Jewish people.