בס"ד

חבורת מהות היהדות
Weekly Newsletter

Vol. 2, Issue 9

פרשת ויקרא
ז ניסן תשפ"ה

Feedback? contact@essenceofjudaism.com

From the Chabura
By: Adam Friedmann

Mapping Morality: AI Alignment and Rav Sa'adia Gaon's Ethical Framework

Two weeks ago we considered the question of AI alignment. Specifically, we dealt with the question of how we can train AI models to align with human needs and values so that using AI results in human flourishing. We noted that several sources in the Jewish tradition indicate that we humans don't have a clear understanding of our own needs, which makes this task difficult if not impossible. This week we'll take a more detailed look at a specific example.

In their paper *What are human values, and how do we align AI to them?* (https://arxiv.org/pdf/2404.10636) researchers from the Meaning Alignment Institute present a novel method for creating an aggregated computer representation of human values. The authors note that people are complex. Any one person's decisions can be motivated by a variety of different values. Furthermore, the way that values interact in different scenarios is also complex. For any given values *a* and *b*, *a* may play a more dominant role than *b* in some cases, and vice-versa. The authors propose a four stage process for creating what they call a "moral graph" that represents different values and their relationships in different contexts.

1. Values are elicited from human participants in conversations with AI chatbots. The chatbot extracts value statements from interlocutors by asking what kinds of values they used to make decisions in a given scenario. For example, someone facing a moral crisis may respond that in the past they turned to religious texts for guidance on how to navigate a similar crisis.
2. The system turns the elicited values into "value cards". These relate a scenario (e.g., a specific type of moral crisis) with a value to use to respond to the scenario (e.g., consulting religious texts).
3. If enough value cards are generated, there will be cases where different people respond by using different values in the same scenarios. For example, some people may respond to a moral crisis by consulting religious texts. Others may ask good friends for advice. Still others may "throw caution to the winds" and do what feels right. In these cases, the system needs to know which value is superior in a given scenario and therefore should be promoted by AI in future interactions. To achieve this, the system contrives stories where people move between values in response to a given scenario. For example, "I used to think that the best way to handle a moral crisis was to choose at random what to do, but I have come to realize that religious texts contain vast wisdom that I can use in these situations."
4. Human participants judge the various value transition stories for a given scenario based on which one presents the "wisest" choice. For whichever story receives the most votes, a connection is made

moving from the value card of the "lesser value" (e.g., handling a moral crisis with random choice) to the value card of the "greater value" (e.g., handling a moral crisis by consulting religious texts).

Using this method a "moral graph" is formed which describes the relative places of different values for a given scenario. AIs can use this graph to promote higher values when interacting with people.

How can we look at this AI alignment system from a Jewish perspective? The proposed system shares many qualities with the template for the virtuous or "good life" that Rav Sa'adia Gaon lays out in the tenth section of his *Emunot Vede'ot*. Rav Sa'adia presents a psychological overview of human drives. He identifies three general drives:

- **Appetitive**: The drive to eat and seek other physical gratification.
- **Impulsive**: The drive to anger, dominance, vengeance, etc.
- **Reason**: The drive to use rational judgement.

In Rav Sa'adia's view, in order to be a moral person, reason needs to be in control. A person then uses rational judgment to decide how to deploy the other two drives in an appropriate manner.

The appetitive and impulsive drives lead people towards certain specific desires. Rav Sa'adia identifies thirteen: 1) abstinence 2) eating and drinking 3) intimate relationships 4) romantic love 5) accumulating money 6) having children 7) material development of the society 8) longevity 9) dominion (eminence, leadership) 10) satisfying the thirst for revenge 11) acquiring wisdom 12) worship 13) rest

According to Rav Sa'adia, focusing solely on one of these desires is unhealthy and immoral. Rather a person must strike the correct balance between them all. This is done by using reason to understand when to satisfy a desire and when to deny it. In this system, the mitzvot serve as a training program through which a person gains the ethical instincts needed for their comprehensive drive to make the right ethical judgements. (See Daniel Rynhold, *An Introduction to Medieval Jewish Philosophy*, 187)

Rav Sa'adia's system can also be used to generate a kind of graph. For a given scenario, a person may respond with deploying or withholding different desires or drives. In each case, we can imagine a hierarchy of possible responses representing better or worse uses of rational faculties. The ideal approach to a scenario is intuited by a person who has fully absorbed the Torah's training. The lines on the graph lead from the worse uses of the drives in a given scenario to the better ones.

The distinction between the moral graph and the one that emerges from *Emunot Ve'deot* is the difference between subjective and objective realities. The moral graph is based on the values that people report to have and their subjective judgements about the wiser course of action in a given situation. Rav Sa'adia's system is based on the objective psychological reality of what drives human behavior and on an instinct for ethical decision-making informed by the Torah. An AI trained on the moral graph would effectively be holding a mirror to humanity and guiding people to behave based on what people think is right, on aggregate. A system trained on Rav Sa'adia's model would, in the ideal case, guide people to make decisions based on *ratzon Hashem*.

Mishnah: A Philosophy of Life
By: Dovid Campbell

Beitzah 1:1 — Creation Eggs Nihilo

This year, Pesach begins immediately after Shabbat, and it therefore seems appropriate to discuss a mishnah that deals with exactly such a case. The opening mishnah of tractate *Beitzah* begins with a deceptively simple question: Can one eat an egg laid on Yom Tov? Beit Shammai permits it; Beit Hillel forbids it. Yet beneath the surface of this halachic debate lies a profound meditation on time, sanctity, and the boundary between human and nonhuman action.

All agree that if the hen was designated solely for egg-laying, both it and its eggs are considered *muktzeh*—set aside and unusable on Yom Tov. So our mishnah must be discussing a unique case in which the chicken itself was designated for consumption. The only remaining issue is *hachana*—a situation in which Yom Tov follows immediately after Shabbat, and yet Shabbat cannot be used to prepare for Yom Tov. This rule protects the spiritual autonomy of these days, reminding us that holy time requires focused attention and cannot serve as a tool for future holiness, just as one does not use Shabbat to prepare for the mundane.

Beit Hillel extends this principle beyond human action, introducing a rabbinic decree that forbids not just deliberate preparation, but even what the Rambam terms *hachana tivit*—natural preparation. In this view, the birth of the egg through the hen's natural processes renders it unfit, as it was "readied" on the Shabbat that preceded Yom Tov, even without human involvement. Beit Shammai disagrees, insisting that the sanctity of the day is only compromised by human effort. Natural processes are not subject to rabbinic extension; they unfold in a different realm of causality.

Fascinatingly, the major commentators use distinct language to describe the emergence of this egg, revealing subtle philosophical leanings. The Rambam calls it *hachana tivit*, natural preparation—a legal category that emphasizes the reality and causal power of nature's rhythms. The *Tiferet Yisrael* chooses the term *memeila*—it happened automatically, of its own accord, highlighting the egg's emergence as something unprompted, beyond deliberate agency. And the Bartenura uses the phrase *biyedei Shamayim*—through the hands of Heaven—underscoring the divine agency that ultimately determines all worldly events.

Each term gestures toward a different way of relating to nature: as a system, as a happening, or as an expression of divine will. And at the heart of this debate is a powerful question: Should the sacred structure of halacha seek to encompass the spontaneous life of nature, or only the conscious acts of man?

Much thanks to Menachem Weinreb for suggesting this week's mishnah! If you'd like to see a specific mishnah explored in this column, please write to us with your suggestion.

Sforno on the Parsha
By: Nochum Spiegel

Humble Beginnings

When one of the partners in a relationship does not live up to their commitment and obligation to the other, a process of reconciliation must occur. In our parsha Hashem reveals the method of elevating a damaged spiritual connection through the offering of *korbanot*.

Continuing the theme which details the repercussions of the *Chet HaEgel* (see Sforno on the Parsha; *Terumah*, *Vayakhel*), Sforno explains (*Kavanot HaTorah*) how the nature of *korbanot* were affected. The previous offerings which we find in the Torah, (*Kayin*, *Hevel*, *Noach*, the *Avot*, at *Har Sinai*) had all been performed on a voluntary basis, not in response to divine command. Post-*Chet HaEgel* there would be added categories of obligatory ones placed on both the individual and community (tamid, musaf, chatas, asham). Detailed processes and procedures were now necessary for *Am Yisrael* to maintain their spiritual stature.

The details and guidelines for *Korbanot* are many, but which is its foundational underpinning? The introductory pasuk states "When any of you presents an offering to Hashem" (*Vayikra* 1:2). Sforno explains (*adam ki yakriv* ***mikem***) the offering must be from yourself (***mikem***), one must present themselves before Hashem with verbal confession and submission, acknowledging their true status before the Creator. Hashem has no desire in the foolish ones who bring offerings without having humbled themselves in preparation. As we explained last week regarding the *Mishkan*, the physical trappings and adornments of our service lack any value if they are devoid of true dedication.

Sforno then turns our attention to the first korban recorded in the Torah, the food offering of Kayin in *parshat Bereishit*. "And to Kayin and his offering Hashem did not turn (find favorable), Kayin was very angry and his face fell". Sforno (*Bereishit* 4:5,9; *Vayikra* 1:2) explains that both Kayin and his offering had separate deficiencies which led to their rejection. The offering was both not of the appropriate species listed in our *parsha*, and Kayin not a fitting offerer. In response to Hashem's question regarding the whereabouts of his murdered brother Hevel, Kayin states that he does not know. Sforno explains that this reveals Kayin's mistaken theological conception that Hashem's knowledge of the actions of man has limitations. He thereby falls into the category of the apostates whose korbanot are not accepted, a law that *Chazal* (*Sifra* to 1:2) will derive from the previously mentioned ***Mikem*** - to exclude the apostate.

The lesson from Kayin is quite telling. The very first offering recorded is rejected by G-d. In place of spiritual connection, there is murder and exile. Lack of a proper understanding of Hashem's relationship with man leads to distance and detachment. Sforno teaches that approaching the service of Hashem without proper preparation can result in a downward spiral of sin. Only those who have truly offered and submitted themselves will attain a true relationship with Him.