Towards use of i* to addressing AI Alignment: a Power and Rationalizing perspective

Daniel Gross

Modal<>AI, HaKeren Hakayement 8, Jerusalem, Israel

Abstract

Recent massive successes of Large Language Models (LLMs) brought the AI Alignment problem to the forefront. This long-standing topic encompasses a range of concerns. For example, to ensure that super-learners are prevented from generating unintended and potentially harmful outputs. And, more broadly, whether AI systems can be imbued with ethics and values that align with users' specific values. Drawing on early works by Brent Flyvbjerg on values and power in human decision-making, we propose viewing alignment as a process where powerful stakeholders strive to actively shape the discourse through rationalizing, thereby obtaining a "license to operate" in the market. To foster a more balanced and equitable discourse we propose employing i* to facilitate a more systematic mapping out and analyzing of stakeholders' rationalizing to revealing unarticulated motivations and goals and alternatives, that were intentionally left hidden in the discourse, but that are crucial to identify supporting or conflicting stakeholders' goals and concerns. By adopting i* we believe that diverse stakeholders in asymmetric, yet interdependent, power relations could be better equipped to analyze power discourses, and thereby achieve more balanced and fair outcomes in the adoption, diffusion and use of powerful AI-based systems within their ecosystem.

Keywords

i*, AI Alignment, Power, Rationalization, Transparency

1. Introduction

The recent tremendous advancements of Large Language Models (LLMs) have demonstrated unparalleled AI capabilities in understanding, responding to, reasoning about, and fulfilling human queries and demands. Presently, virtually every major company is actively integrating LLMs into their technological infrastructure.

With such breathtaking developments accelerating at an unprecedented pace across industries, key questions of AI safety and Alignment [3, 4, 5] have come to the forefront.

Concerns of AI safety, and the related topic of AI alignment, is already more than a decade old and encompasses a range of concerns. From technical concerns: whether super-learners such as LLMs that traverse vast state spaces can be prevented from arriving at unintended and even harmful outputs [4]. And, more broadly, whether AI systems can be imbued with the kind of ethics and values that underpin the human users these systems come to serve [5].

We posit that such concerns of technical and value alignments are indeed critical to the successful and safe operation of AI system within human society. However, we recognize another significant aspect of alignment: the alignment of interests among stakeholders involved in the creation, deployment, use of, and that are impacted by AI-based solutions introduced to the market.

With the emergence of today's powerful AI technologies, that are about to change everything, there are vocal voices of powerful stakeholders, who have a lot to gain from adoption of AI, and that seek broad legitimacy in operating such systems in markets.

The 16th International iStar Workshop, September 03–04, 2023, Hannover, Germany
EMAIL: grossd18@mail.com (D. Gross)
ORCID: 0000-0003-2381-8975

Following the early works of Flyvbjerg [1,2], we propose viewing AI alignment through the lens of power and rationalizing. Flyvbjerg states: "[p]ower determines what counts as knowledge, what kind of interpretation attains authority as the dominant interpretation. Power procures the knowledge which supports its purpose, while it ignores or suppresses that knowledge which does not serve it." [1]. Furthermore, "the relationship between knowledge and power are decisive, if one seeks to understand the kinds of processes affecting the dynamics of politics, administration, and planning" [1].

Flyvbjerg presents ten propositions that capture key aspects of power and knowledge, and related rationalizing processes. The ten propositions are derived from studies of power in the field and include propositions such as "Power defines reality", "Rationality is context-dependent, the context of rationality is power, and power blurs the dividing line between rationality and rationalization", and the like.

Based on these principles Flyvbjerg posits that powerful stakeholders exert their power by defining reality, which includes deciding what knowledge to make available, and what rationalizations to thereby offer to influence outcomes favorably to their goals.

Flyvbjerg further emphasizes that rationality, the establishing of the normative, "what should be done", does not survive when it meets "what is actually done", which is what power determines in practice [1]. Instead of viewing rationale discourses, as a means for dissolving power, Flyvbjerg argues for a Realrationalität, a rationalizing that takes power as it is exerted in practice, on the ground, into account, with an aim to regulate power and domination.

We see the i* modeling approach in particular well suited to offer support for Realrationalität. Unlike mechanistic modeling methods common in the software requirements and engineering domain, the i* modeling approach looks to bring a social understanding to rationalizing processes by adopting a social ontology for its main modeling constructs [9].

Furthermore, the paradigm underpinning the i* modeling approach, and the social semantics afforded to its constructs, in particular, looks to deal with inherent inability to know and control the social world [9].

The i* modeling method therefore looks to provide analysis tools to externally characterize actors using rationalized intentionality and interdependencies. External characterizations are then a means for external observers to more systematically explore hidden (semi-opaque) motivations, knowledge and intents of actors pursuing their interest.

Adopting Flyvbjerg, we would say that knowledge that could be shared is however left unarticulated by actors as a key part of exerting power. We see overcoming such attempted obfuscation by using i* modeling and analysis capabilities as part of a Realrationalität alignment effort of stakeholder interests, to ensure equitable and fair outcomes.

The next section presents approaches of how power has been described and analyzed in relation to technology, and contrasts these with rationalizing as explained by Flyvbjerg, and proposed here. Section three illustrates the application of power through constructed rationalizing taken from a study performed by Flyvbjerg. Section four suggests how i* could be used as a tool to overcome such constructed rationalizing by looking to systematically broadening the discourse space.

Finally, section five concludes by pointing to key challenges in the use of a modeling technique in power discourse settings, and future work.

2. Describing and Analyzing Power

Milne and Maiden [6] explain that power is multifaceted and involves several dimensions, including a) the ability to directly influence the behavior of another party, b) to define reality for another party, such as set an agenda, and c) to determine values and norms and ideologies that underpin decision making.

Milne and Maiden then propose a modeling technique based on social networks whereby directed links between stakeholder nodes are labeled to indicate what kind of power a source party can exert on a target party. Types of power include legitimacy through formal authority, reward, coerce, expertise, and power through personal characteristics.

Milne and Maiden further discuss the limitation of the i*'s strategic dependency model to capture power relationships and suggest augmenting strategic dependencies links with their labeling.

Alves, Valença and Franch [7] examine power relationships in software ecosystems and propose guidelines on how companies can identify and apply power. More specifically, Alves, Valença and Franch focus on a power capability, a resource controlled by a party needed by other parties, thereby enabling the "power source" party to exert power and control over other parties. Cultivating such a resource, and relationships with parties needing such resources, is creation and use of power.

Approaches such as above focus on a relational characterization of power, that allows labeling and then analyzing power relationships in terms of graph properties.

Offering an important step towards alignment, Schlichter [8] identifies key needs to systematically address ethics and alignment in the development and use of AI. Drawing from works of Flyvberg [1] Schlichter offers a catalogue of questions to help establish five organizational maturity levels, with level five offering evidence that organizational processes are established to seek value alignment across stakeholder in the ecosystem the organization operates in.

Based on the work by Flyvbjerg [1,2] and in line with Schlichter [8], we posit that power is exerted at the discourse contents level: the knowledge and arguments that a power party offer to rationalize decisions they promote. Furthermore, powerful parties may also deceptively withhold knowledge to support their rationalizing and to get buy-in from 3rd parties whose legitimacy they seek in the market, say, such as from customers and regulators.

We therefore see as a key need to systematically and critically explore rationalizing arguments, looking to identify unarticulated knowledge and intents.

Seen from this vantage point, we posit that key value can be derived from applying the i* modeling approach to systematically explore the open landscape of stakeholders, and alternatives, including scopes of decision-making, while putting a strong methodological emphasis on the underpinning social ontological modeling assumptions of i* -- the need to externally characterize unknowable, and uncontrollable stakeholders – and to further augmenting, as needed, i*'s social constructs to deal with power relationships between actors.

3. Example of exertion of power through controlling knowledge and rationalizing

Flyvbjerg [1, 2] explores case studies in urban planning, shedding light on power and rationalizing at a Danish town. In this instance, a prominent stakeholder with significant power, namely the mayor, advocated, along with his planners, for the construction of a bus terminal near the transportation system.

The mayors, and his planners, argument centered around the objective need of reducing transfer time for travelers, presenting a seemingly compelling rationale to relevant stakeholders involved in this discourse, such as environmentalists, businesses, and the city architect's office.

However, concealed within this reasoning was the fact that the transportation system itself had been intentionally planned to justify the location of the bus terminal, in alignment with the mayor and his planners' desires.

While providing stakeholders with a seemingly plausible justification for the chosen bus terminal location, the circularity of the decision-making process, the broader planning process and real intents, remained hidden.

Flyvbjerg explains that "[i]n real terms, by playing games of power covered up as technical reasoning, the mayor and the planners got what they wanted – a monument to themselves – despite rampant protest …"

Consequently, real concerns, which eventually materialized, regarding the bus terminal's location, such as increased traffic accidents, elevated levels of air and noise pollution, and physical blight, were rationally traded-off for speed of transfer, while also keeping the discourse scope focused on the terminal's location only, and not the broader planning process.

To expose such rationalizing in the service of power, Flyvbjerg calls for a study of realrationalität (real-world rationalizing) "to depict what the lived, as opposed to the ideal, world of modern democracy and planning looks like and how it operates …".

It's a method that goes after the details, and that looks to unearth hidden motivations and choices.

4. Towards the use of i* for analysis power and rationalizing

We believe that i* [9] can serve as a modeling approach at the intentional level for exploring rationalizing arguments presented by stakeholders in positions of power. The use of strategic dependency models allows for the examination of the types of relationships sought by powerful stakeholders, while also providing a systematic means for exploring viable alternative relationships.

Strategic rationale models support the exploration of possible hidden internal motivations and intents that are linked to, and explain, the external dependencies. The hierarchical structure of rationale models enables considering the provided rationales as starting points, to explore higher-level intents (the 'why') and possible alternatives (the 'how else'), which could then lead to alternative dependency configurations.

The open nature of the model enables a systematic broadening the scope of discourse, including, to alternative that have not been included in the discourse.

To illustrate, let us consider again the bus terminal location being justified by the need for fast transfers. In this case, one could use the strategic rational structures to start question the underlying reasons for the chosen locations of the transportation system, thereby uncovering potential hidden rationales, or lack thereof, that necessitate placing the transportation system at its location, in the first place. By moving up the rational hierarchy structures the topic “bus station” could be expanded to transportation system, and beyond, and alternatives and rationales for the overall design explored, leading to questions that expand the scope of discourse and thereby making it more difficult to obfuscate the overall design of the transport system including its locations.

An i* modeler would need to be critically aware that what may be presented as a constraint, the transport system locations, modeled as a claim softgoal, which then would not be part of the alternative exploration, may in fact be a hidden choice, better modeled as a task, or even an actor, and open for further exploration, evaluation and negotiation.

By engaging in such inquiry, it becomes possible to bring additional goals and choices of relevant stakeholders in more systematic manner into the conversation, such as reducing traffic and accidents, and pollution, and the rationales for the locations of the transport system as a whole in general.

Given the conceptual fit of i* with an exploration of rationalizing in social settings, we belief that additional empirical research that examine the tactics employed by powerful entities to rationalize and construct reality would lead to fruitful insights into further expressivity needs of i*, as well as the development of further key analysis methods specifically tailored for analyzing power discourses and alignment.

5. Discussion & Conclusion

The recent advancements in AI, particularly LLMs, are attracting influential stakeholders who have a substantial stake in the adoption of these AI technologies. These stakeholders are actively engaged in advocating for the remarkable value of emerging AI solutions, while also justifying the transformative impact AI will have on the workforce, and society at large. As a result, there has been a remarkable increase in the creation and deployment of AI-based solutions across various aspects of our lives, which will see further tremendous acceleration in the future.

Given the magnitude of these developments, there is a critical need for the development of robust and easy to use analysis tools of power-driven rationalizing. The i* modeling approach is well positioned to offer along with additional tools such as Schlichter’s ethical maturity model, could play a pivotal role in facilitating the process of alignment of interests between stakeholders, addressing a key requirement in the field of AI.

Clearly, the challenge of intentionally obfuscating knowledge is a significant one, in particular in a power setting, and the mere use of a modeling technique may not make a difference. Yet, there is significant value in making verbal and written, argumentation, visible in models, and thereby enabling a more systematic analysis of the problem domain, and to provide tools for communicating knowledge and decision-making, which is a key value of modeling tools in general.

An i* based modeling and analysis at the intentional level could also make a significant contribution to establishing AI systems alignment with human values, by making visible and amendable for machine learning stakeholder motivation and goals through strategic dependencies and rationales.

The use of the i* modeling method could therefore also provide significant value for other aspects of alignment when deploying AI based systems in social settings.

We therefore advocate for further advancements in, and diffusion of, i* to support its adoption and use in the context of AI alignment, and the analysis of power and rationalizing. The requirements engineering (RE) community is uniquely positioned to make significant contributions to more equitable aligned AI based systems during this significant inflection point in time.

6. References

[1] Flybjerg, Bent "Rationality and Power". In Scott Campbell and Susan S. Fainstein, eds. Readings in Planning Theory, 3rd edition, Oxford: Blackwell, pp. 318-329 1998.

[2] Flybjerg, Brent “Aristotle, Foucault and progressive phronesis: an outline of an applied ethics for sustainable development”. In E.R. Winkler and J.R. Coombs, eds. Applied ethics: a reader. Cambridge, MA

[3] Taylor, Jessica, Eliezer Yudkowsky, Patrick LaVictoire and Andrew Critch. “Alignment for Advanced Machine Learning Systems.” Ethics of Artificial Intelligence 2020

[4] Shalev-Shwartz, S., Shammah, S., & Shashua, A. ”On the Ethics of Building AI in a Responsible Manner”. ArXiv, abs/2004.04644 2020.

[5] Han, S., Kelly, E., Nikou, S., & Svee, E. O. ”Aligning artificial intelligence with human values: reflections from a phenomenological perspective”. AI and Society, 37(4). https://doi.org/10.1007/s00146-021-01247-4 2022.

[6] Milne, A., Maiden, N. ”Power and politics in requirements engineering: embracing the dark side?”. Requirements Eng 17, 83–98. https://doi.org/10.1007/s00766-012-0151-6 2012

[7] C. Alves, G. Valença and X. Franch, "Exercising Power in Software Ecosystems," in IEEE Software, vol. 36, no. 3, pp. 50-54, May-June 2019, doi: 10.1109/MS.2018.290101618.

[8] "Schlichter, J. E-AIMM 1.0; Advenæ, September 11th. — Advenæ. http://advenae.ai/e-aimm." (2019).

[9] Yu, E.S. (2009). Social Modeling and i* . In: Borgida, A.T., Chaudhri, V.K., Giorgini, P., Yu, E.S. (eds) Conceptual Modeling: Foundations and Applications. Lecture Notes in Computer Science, vol 5600. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-02463-4_7