Genies, lawyers, and smart-asses: Extending proxy failures to intentional misunderstandings

Tomer D. Ullman[1] & Sophie Bridgers[1,2]
[1]Harvard University
[2]Massachusetts Institute of Technology

Citation: Ullman, T. D., & Bridgers, S. (2024). Genies, lawyers, and smart-asses: Extending proxy failures to intentional misunderstandings. Behavioral and Brain Sciences, 47, e86.

ABSTRACT

We propose that the logic of genie – an agent that exploits an ambiguous request to intentionally misunderstand a stated goal – underlies a common and consequential phenomenon, well within what is currently called proxy failures. We argue that such intentional misunderstandings are not covered by the current proposed framework for proxy failures, and suggest to expand it.

MAIN

Making your way through busy market stalls, you chance upon an antique lamp. As you brush the dust off to inspect it, a genie springs out in a cloud of colorful smoke.

“One wish, no more, no less,” says the genie.

“Make me rich!” you reply, and immediately alarm bells go off in your head. Your mind floods with tragic tales of people who got what they asked for.

“Don’t kill my parents so I inherit their money,” you hasten to add. “Actually, don’t kill anyone. Also I don’t want stolen mafia money. Or any kind of stolen money. Make that ‘no crimes’. And don’t make it that I turn things into gold, I just want money. Real money. And don’t make everyone poor so I’m rich by comparison, and…” you trail off, thinking it through.

Eventually, you place the lamp down carefully and say, “You know what? Forget I even asked.”

Compared to the examples of proxy failure in John et al., our genie example seems fanciful. But, we propose the logic of the genie – an agent that exploits ambiguous requests to intentionally misunderstand the stated goal – underlies a phenomenon that is (1) well within ‘proxy failures’, and (2) common and consequential, but (3) not covered by the current framework of John et al. Our point is not that John et al.’s framework is wrong; we find it both enlightening and useful.

Rather, we suggest that many important situations that seem to fall under the notion of proxy failure require expanding their framework. Our argument is not a “no”, but a “yes, and”.

To start, the dynamics of intentionally misunderstanding requests follow the logic of many of the examples John et al. use when introducing the problem of interest. A tenant asked by their landlord to “do some weeding", who then pulls out three weeds and calls it a day, is acting in line with the terms used by John et al.: A regulator (landlord) with a goal (cleared yard) conveys the goal to an agent (tenant), but uses language that doesn’t match the goal directly, after which the agent engages in “hacking” or “gaming”.

Intentional misunderstandings are common and important. They show up in history (Scott, 1985), fables and art (Da Silva & Tehrani, 2016; Uther, 2004), childhood (Opie & Opie, 2001), and interpersonal conflict (Bridgers et al., 2023). Such letter-vs-spirit of the law concerns are also often discussed in the legal realm (Hannikainen et al., 2022; Isenbergh, 1982; Katz, 2010). But such concerns are not with scalar proxies standing in for a true but unknowable goal. Consider a lawyerly child watching videos on their tablet who is told, “time to put the tablet down," and proceeds to place the tablet on a table, only to keep watching their videos. Such a child is not optimizing a scalar reward conveyed by a parent who cannot convey some complex goal and resorts to a proxy. The parent was being quite clear, and the child was being quite a smart-ass.

If we accept the above, then intentional misunderstandings pose a challenge for the framework of John et al. The framework supposes that a regulator has a difficult-to-convey goal, and instead gives an agent a different goal. But in many current models of human communication, concepts (including goals) are hidden variables, conveyed indirectly through ambiguous utterances (Goodman and Frank, 2016). This is true for any goal, including proxies. The process of recovering meaning from ambiguous utterances is usually so transparent that people don't even notice it unless it breaks down, e.g. when hijacked intentionally. To see this, take the Hanoi rats (please): The original goal of killing all the rats is unobserved, but can be easily recovered from the utterance “kill all the rats”. The utterance “bring me rat tails” is not a proxy goal, it is an utterance, which can be used to derive the original goal, and is likely understood by people to mean the original goal. True, the proxy utterance can be intentionally misunderstood, but so can the original utterance – there is nothing inherently special about proxy utterances in terms of clarity from the standpoint of a theory of communication, and likewise there is nothing inherently special about a proxy goal in terms of observability.

Our differing analysis for intentional misunderstandings is important for at least two reasons: First, it moves the focus away from illegibility and prediction, especially in communicating goals to machines. People who supposedly convey a ‘proxy goal’ to a machine do not experience failure because they can’t predict all the ways in which their proxy goal might have unintended consequences. Rather, they experience failure because they didn't even realize they were conveying a different goal in the first place (many of the examples in Krakovna, 2020 are like this).

An engineer evaluating a loss function infers the goal behind it, because that is how human communication works, but most machines currently aren’t built to run an inference process from a loss function to an intended goal. This brings us to a second reason the differing analysis is important: it suggests currently unexplored remedies, at least for some cases. Telling a genie (or child, or lawyer) a goal, and then tacking on a long list of caveats will not stop them if they are determined to misunderstand: Every caveat is an opportunity for another loophole (in line with the 'proxy treadmill'). By contrast, highlighting common-ground in order to specifically rule out loopholes is useful, e.g. telling a child "can you do your homework?" and following it with "you know what I mean" to avoid the tired “yes, I *can*”. Highlighting common-ground would not be useful for a machine that is optimizing a given loss function rather than engaging in human-like communication.

And what of our genie? They forgot you even asked, just like you wanted. And so, they offer you one wish. No more, no less.

ACKNOWLEDGEMENTS

We thank GU for unintentionally providing some of the examples of intentional misunderstandings. TDU is supported by a Jacobs Foundation Fellowship. This commentary makes reference to work supported by an MIT Simons Center for the Social Brain Postdoctoral Fellowship (SB), and an NSF Science of Learning and Augmented Intelligence Grant 2118103 (TDU, SB).

REFERENCES

1. Bridgers, S. E. C., Taliaferro, M., Parece, K., Schulz, L., & Ullman, T. (2023). Loopholes: A window into value alignment and the communication of meaning.
2. Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818-829.
3. Hannikainen, I.R., Tobia, K.P., de Almeida, G.d.F., Struchiner, N., Kneer, M., Bystranowski, P., Dranseika, V., Strohmaier, N., Bensinger, S., Dolinina, K., et al.: Coordination and expertise foster legal textualism. Proceedings of the National Academy of Sciences 119(44), 2206531119 (2022)
4. Isenbergh, J.: Musings on Form and Substance in Taxation. HeinOnline (1982)
5. Katz, L.: A theory of loopholes. The Journal of Legal Studies 39(1), 1–31 (2010)
6. Krakovna, V.: Specification gaming examples in AI - master list. http://bit.ly/kravokna_examples_list Accessed: 2020-12-28 (2020)
7. Opie, I.A., Opie, P.: The lore and language of schoolchildren. New York Review of Books (2001)

8. Scott, J.C.: Weapons of the weak: Everyday forms of peasant resistance. Yale University Press (1985)
9. Uther, H.-J.: The types of international folktales–a classification and bibliography. Suomalainen Tiedeakatemia Academia Scientiarum Fennica Exchange Centre (2004)