Responsible bots: 10 guidelines for developers of conversational AI

More people are using bots in their everyday lives, whether it's to get a quick answer to a customer service problem or to help people out with things like managing their calendars, checking the weather or ordering pizza. Bots, or more generally, conversational AI, have the ability to help people achieve more, and we are only starting to see their potential to augment what we can do.

In order for people and society to realize the full potential of bots, they need to be designed in such a way that they earn the trust of others. These guidelines are aimed at helping you to design a bot that builds trust in the company and service that the bot represents. These guidelines are not intended as legal advice and you should separately ensure that your bot complies with the fast-paced developments in the law in this area. Also, in designing your bot, you should consider a broad set of responsibilities you have when developing any data-centric AI system, including ethics, privacy, security, safety, inclusion, transparency and accountability. See, for example, Microsoft's six principles for the responsible development of AI published in the January 2018 book, The Future Computed.

These guidelines are just that — guidelines, and for the most part not hard-and-fast rules. They are most relevant to bots that may affect people in *consequential ways* — such as helping people to navigate information relating to employment, finances, health or the like. You should use your best judgment when applying these guidelines, always with a view toward the ultimate question of whether your design will deliver an experience that end users appreciate, in a manner that builds their trust in your company and services. These are v.1.0 guidelines, so we fully expect that they will be revised over time in response to your feedback and our own experiences.

Guidelines

1. Articulate the purpose of your bot and take special care if your bot will support consequential use cases.

The purpose of your bot is central to ethical design, and ethical design is particularly important when it is anticipated that a consequential use will be served by the bot you are developing. Consequential use cases include access to services such as healthcare, education, employment, financing or other services that, if denied, would have meaningful and significant impact on an individual's daily life.

- **Before beginning design work, carefully specify how your bot will benefit the user or the entity deploying the bot.** If the bot is likely to affect the well-being of the user, such as providing access to a consequential service like healthcare, attention to these guidelines will be especially important. Be sure to pause to research, learn and deliberate on the impact of the bot on people's lives. When in doubt, seek guidance.
- **Assess whether the bot's intended purpose can be performed responsibly.** Some purposes may inherently require human judgment, empathy and expertise or a very high degree of reliability and accuracy, e.g., healthcare diagnosis or financial planning. Be sure to consider the nature and type of errors in the performance of the bot and their cost to users. Consider if you have access to relevant expertise in the domain in which your bot would operate.
- **Develop metrics to assess user satisfaction.** Metrics for your bot should cover not only whether the user feels that the bot served its intended purpose, but also the user's sense of well-being and comfort while using the bot.

2. Be transparent about the fact that you use bots as part of your product or service.

Users are more likely to trust a company that is transparent and forthcoming about its use of bot technology, and a bot is more likely to be trusted if users understand that the bot is working to serve their needs and is clear about its limitations.

- **It should be apparent to the user that they are not having an interaction with another person.** Since designers might endow their bots with "personality" and natural language capabilities, it is important to convey to users that they are not interacting with another person and some aspects of their interaction are being performed by a bot. There are variety of design choices that can be made to accomplish this that do not degrade the user experience.
- **Establish how the bot can help and the limitations associated with its use.** Users are more likely to find a bot to be trustworthy if the bot sets reasonable expectations for what it can do and what it does not do well. Users should be able to easily find information about the limitations of the bot, including the possibility of errors and the consequences that can flow from such errors. For users who wish to "learn more," you should offer a more detailed explanation of the purpose and operation of the bot.

3. Ensure a seamless hand-off to a human where the human-bot exchange leads to interactions that exceed the bot's competence.

If your bot will engage people in interactions that may require human judgment, provide a means or ready access to a human moderator.

- **Respect individual engagement preferences, particularly if your bot deals in consequential matters.** Bots designed for use in consequential matters should incorporate the ability to transfer an engagement to a human moderator as soon as the user asks, otherwise indicates, or if the bot recognizes (e.g., through sentiment analysis) that the user is dissatisfied. If users feel trapped or alienated by a bot, they will quickly lose trust in the technology and in the company that has deployed it.

4. Design your bot so that it respects relevant cultural norms and guards against misuse.

Since bots may have human-like personas, it is especially important that they interact respectfully and safely with users and have built-in safeguards and protocols to handle misuse and abuse.

- **Limit the surface area for norms violations where possible.** Every bot should be designed to follow a specific set of values and cultural norms. To reduce the possibility of conflicting with those values and cultural norms, limit the surface area for norms violations. For example, if your bot is designed to take pizza orders, limit it to that purpose only, so that it does not engage on topics such as race, gender, religion, politics and the like.
- **Where appropriate, point to a relevant "code of conduct" for users.** Consider whether your bot should be subject to a user code of conduct (from your organization or the entity deploying the bot) that, for example, includes prohibitions on hate speech, bullying and threatening others, and provides appropriate notice to the user of any code of conduct.
- **Apply machine learning techniques and keyword filtering mechanisms to enable your bot to detect and — critically — respond appropriately to sensitive or offensive input from users.** Deploy a two-way filtering mechanism with a customizable threshold of tolerance for what your bot takes in from users, as well as what your bot says in response. In most cases, we

recommend the bots simply steer clear of controversial subjects (especially hate speech). Open domain conversations are considered high-risk because they require significant investment in both content operations and social media monitoring capabilities and must be maintained 24/7 with bugfix service level agreements. You should leverage products that include offensive text classifiers, such as the Microsoft Bot Framework, , to protect your bot from abuse if it engages in open domain conversations. Sensitive categories include adult content, extremism, drugs, alcohol and tobacco, profanity, vulgarity, harassment, bullying, violence and gore, and hate speech (relating, for example, to ethnicity or race, gender identity or sexuality, religion, or people with disabilities). Public-facing bot APIs should also be reviewed to assess whether they could be used by people outside your organization to create a bot that would engage in hate speech or otherwise reflect poorly on your organization.

5. Ensure your bot is reliable.

Ensure that your bot is sufficiently reliable for the function it aims to perform, and always take into account that since AI systems are probabilistic they will not always provide the correct answer.

- **Establish reliability metrics and review them periodically.** Consider what questions your bot needs to answer and rigorously test its performance and ongoing effectiveness. Because the performance of AI-based bot systems may vary from development to implementation, and over time as the bot is rolled out to new users and in new contexts, it is important to continually monitor reliability. Reliability signals can be developed to help drive decisions about when to pass the baton to a human, or when a bot should announce that it cannot perform the requested function reliably. If an AI-based bot system can determine that it has made a mistake, that fact should be communicated to the user.

- **Be transparent about bot reliability.** Particularly for bots operating in sensitive domains, you should make available information concerning the reliability of the bot, such as summaries of general statistical performance, performance under particular circumstances, or in the context of specific examples.

- **Build traceability capabilities into your bot.** When something goes wrong with your bot during a high-value interaction, it is important to have traceability (monitoring and auditing), for example through Microsoft Azure Application Insights, in order to troubleshoot the issue. For more information on Application Insights, refer to: https://azure.microsoft.com/en-us/services/application-insights/.

- **Provide a feedback mechanism.** Users will feel more comfortable with bots if they can provide feedback on their operation (and feedback is essential in any event, as with all product development work). Bots should actively ask for feedback. Set expectations as to whether the user will get any response to feedback provided.

- **For sensitive uses, obtain domain expertise.** If you are building a bot to deliver services in areas such as health, employment, finance or law enforcement, ensure that you obtain and take account of input from experts in these areas as you design and deploy your bot.

6. Ensure your bot treats people fairly.

The possibility that AI-based systems will perpetuate existing societal biases, or introduce new biases, is one of the top concerns identified by the AI community relating to the rapid deployment of AI. Development teams must be committed to ensuring that their bots treat all people fairly.

- **Systematically assess the data used for training your bot.** Systematically assess the data used for training your bot to ensure that it has appropriate representativeness and quality, and take steps to understand the lineage and relevant attributes of the training data. As bias detection tools become more broadly available, adopt them as an additional means to ensure the fairness of your bot and make such tools available for customer use and adoption.
- **Strive for diversity amongst your development team.** Employing a diverse team focused on the design, development and testing of bot technology will help ensure that your bot operates fairly.

7. Ensure your bot respects user privacy.

Privacy considerations are especially important for bots. While the Microsoft Bot Framework does not store session state, you may be designing and deploying authenticated bots in personal settings (like hospitals) where bots will learn a great deal about users. People may also share more information about themselves than they would if they thought they were interacting with a person. And, of course, bots can remember everything. All of this (plus legal requirements) makes it especially important that you design bots from the ground up with a view toward respecting user privacy. This includes giving users sufficient transparency into bots' data collection and use, including how the bot functions, and what types of controls the bot offers users over their personal data.

- **Inform users up front about the data that is collected and how it is used and obtain their consent beforehand.** Provide easy access to a valid privacy statement and applicable service agreement and include a "profile page" for users to obtain information about the bot with links to relevant privacy and legal information. Making this information available and easily accessible in the "first run" experience is highly recommended.

- **Collect no more personal data than you need, limit access to it and store it for no longer than needed.** Collect only the personal data that is essential for your bot to operate effectively. If your bot will share data (such as with another bot), be sure only to share the minimum amount of user data necessary in order to compete the requested function on behalf of the user. If you enable access by other agents to your bot's user data, do so only for the minimum time necessary in order to compete the requested function. Always give users the opportunity to choose which agents your bot will share data with and what data is suitable for sharing. Consider whether you can purge stored user data from time to time while still enabling your bot to learn. Shorter retention periods minimize security risks for users and will help to position your bot as privacy-friendly.

- **Provide privacy-protecting user controls.** For bots that store personal information, such as authenticated bots, consider providing an easy-to-find "Show me all you know about me" button, and similar buttons to "Forget my last interaction," "Delete all you know about me," and so forth. In some cases, such buttons may be legally required.

- **Obtain legal and privacy review.** The privacy aspects of bot design are subject to important and increasingly stringent legal requirements. Be sure to obtain both a legal and a privacy review of your bot's privacy practices through the appropriate channels in your organization.

8. Ensure your bot handles data securely.

Users have every right to expect that their data will be handled securely. Follow security best practices that are appropriate for the type of data your bot will be handling.

- **Establish secure development and secure operations foundations.** Traditional secure software foundations are critical. As with any AI system, your bot should ensure proper authentication, separation of duty, input validation and mitigations for denial-of-service attacks.

- **Your bot should be resilient.** Design your bot to identify abnormal behaviors and prevent manipulation. Pinpoint "users" (who could in fact be malicious bots) who deviate from norms established by large clusters of other users — such as users who seem to respond too fast, don't sleep, or trigger parts of your bot code paths that other users do not.

- **Ensure the integrity of your training data.** All AI systems must be able to distinguish between maliciously introduced data (which must be purged) and data that is merely rare, yet valid and potentially important. This is particularly critical for bots which employ automatic or supervised learning techniques.

- **Obtain security review.** If available, work with the appropriate security team within your organization to conduct a security review on your bot and supporting services. Given the close relationship of security and privacy in this space, a joint security/privacy review is recommended to ensure the best depth and breadth of coverage.

9. Ensure your bot is accessible.

Bots can benefit everyone, but only if they are designed to be inclusive and accessible to people of all abilities. Microsoft's mission to empower every person to achieve more includes ensuring that new technology interfaces can be used by people with disabilities, including users of assistive technology.

- **If you are developing a bot, consider how your bot complies with commonly used accessibility standards, such as WCAG 2.0 AA, and U.S. Section 508 and EN 301 549 standards.** Customers with disabilities should be able to use your bot as effectively as those without disabilities. Complying with the international web accessibility standard WCAG 2.0 AA (codified as ISO 40500:2012) and U.S. and European procurement standards will help enable users who rely on screen readers, navigate UI using only keyboard, are hard of hearing, require color contrast or cannot distinguish between colors, to use your bot. Many of these requirements carry dependencies on the conversational canvas.

- **Have people with disabilities test your bots.** In addition to complying with accessibility standards, getting feedback from users with disabilities on your bot prior to launch will help determine whether the bot can be used as intended by the broadest possible audience.

- **Design bots to respect the full range of human abilities.** Use Microsoft's Inclusive Design toolkit to design bots which recognize exclusion, learn from diversity and solve for ability constraints.

10. Accept responsibility.

We are a long way away from bots that can truly act autonomously, if that day will ever come. Humans are accountable for the operation of bots.

- **Developers are accountable for the bots they deploy.** If you are developing a bot that your organization will deploy, you should recognize that you are fully responsible for its operation and how it affects people. If you are designing a bot to be deployed by a third party, come to a shared understanding with them of who is ultimately responsible for the bot and document that understanding.

Microsoft Corporation
Last revised: November 2018