

Author(s): Norman Mooradian, Ph.D.

## Conversational Agents and Personal Privacy Harms Case Study

### Abstract:

This fictional case study examines the question of whether a personal conversational agent/advisor called iSoph, encoding extensive personal information and having a fluent natural language interface, may raise privacy harms that are normally thought to attend to personal relationships, as opposed to harms associated with institutional databases and big data analytics. The case stipulates that (i) iSoph collects extensive personal data about its users from multiple, multimodal sources, (ii) can make inferences from this data in combination with its models, but (ii) cannot share information with its developer or any third party. It is shown that despite condition (iii), iSoph raises privacy risks. These privacy risks are of the type associated with personal relationships and direct observation. iSoph raises these risks because, as an advanced conversational agent, it is able to evoke anthropomorphizing responses from its users in ways that they are not fully conscious of or able to control.

### Agenda:

<b>Case Description 1</b>	<b>2</b>
<b>Questions 2</b>	<b>3</b>
<b>Exercises 3</b>	<b>3</b>
<b>Applying the Principles of AI Ethics 4</b>	<b>3</b>
<b>Normative Theories 5</b>	<b>4</b>
<b>Expert Analysis 6</b>	<b>4</b>
Description of AI Principles and Privacy Harms 6.1	5
Ethics of Simulating Sentience 6.2	6
Interpersonal Privacy Harms and Conversational Agents 6.3	7
Privacy Benefits 6.4	8
Ethical Design 6.5	9
<b>Student Reflection 7</b>	<b>9</b>

### Author(s):

- Norman Mooradian, Ph.D.
- School of information, College of Information, Data, and Society, San José State University, One Washington Square, San José, CA 95192-0
- ✉ [norman.mooradian@sjsu.edu](mailto:norman.mooradian@sjsu.edu), ✉ [Dr. Norman Mooradian Faculty Profile \(sjsu.edu\)](#)

## Case Description 1

The following scenario is envisioned. A personalizable conversational agent system named “iSoph” is developed and marketed by an artificial intelligence developer with the name SophicArts. The purpose of iSoph is to serve as a personal advisor to its users that can advise on a wide range of matters. These include mundane matters such as entertainment choices, as well as more impactful issues such as career choices, health issues, personal relationships, religious and ethical commitments, and general life plans.

iSoph is architected using a full range of natural language processing (NLP) technologies, such as large language models (LLMs), speech-to-text and text-to-speech, as well as other AI technologies such as recommendation engines, as well as knowledgebases and reasoning engines. A large store of general factual knowledge is encoded in the system based on multiple training methods using vast amounts of data available on the internet, as well as specialized domains such a health research database (e.g., PubMed). (Here, factual knowledge is understood broadly to include social facts, normative statements, and value judgements, etc.) iSoph’s capabilities include a natural language interface that allows users to speak to the system or type in a chat box and receive answers in audio or textual form, as desired. The system’s verbal output is generally indistinguishable from human speakers within normal domains (e.g., excluding mathematical questions). Further, iSoph does not need to wait for input from its users. It can comment based on the user’s interactions with his/her computer system and behaviors captured by video.

The iSoph system is run on a dedicated, specialized computer that allows downloads/updates from a secure, cloud-based site, but it does not allow uploads or transmissions of data back to the cloud site. Downloads include program and model<sup>1</sup> updates. The system attaches to a personal computer via a high bandwidth data connection in order to monitor and train on data transactions and stored data on the personal computer. These include browsing (web sites, social media), reading/viewing (web pages, video), input (writing documents, messages), and content stores. As with its connection to the SophicArts cloud site, the iSoph system does not output any data to the attached PC/Notebook. It simply uses it as a unidirectional data source. iSoph can also attach to mobile devices to allow it to scan app histories and collect data, or it can download synchronized device data from the SophicArts site. Finally, iSoph uses a set of cameras directly attached to it in order to observe its users’ behavior.

iSoph’s ability to converse with its users about personal matters is based on AI technology built into the system. This technology uses the data and models available on the SophicArts cloud, but more importantly, it generates its own models and creates a local knowledgebase using AI algorithms. To do this, it uses data collected through a series of forms presented to the user upon setting up the system. These forms solicit biographical information (education, health information, etc.) The forms are periodically updated. iSoph also uses all data traffic on the attached PC/Notepad as described above and video/audio feeds from its cameras. These sources of data account for the lion’s share of information iSoph uses to train its local models and create its personalized knowledgebase. **The information on the iSoph system is completely secured. It is not shared outside of the system and its interface.** Its users are authenticated via biometrics, dual authentication, and challenges. The data is fully encrypted.

The target (ideal) user is an early adopter and heavy user of computer technologies, especially for social computing, but not a computer scientist, data scientist, programmer, etc.

## Summary Points

---

<sup>1</sup> An AI model is the output of machine learning algorithms. Algorithms train on data and learn a model that can be thought of as representation of how a set of variables (features) relate to a target variable. Once learned, a model can be applied to inputs in order to classify / predict features of the things described by the inputs.

1. The iSoph system captures large amounts of its users’ personal data.
2. It uses this information to generate inferences to new personal information, some of it highly sensitive, as well as potentially surprising to its users.
3. It provides information in the form of conversation (answering questions, making comments, etc.) in a manner practically indistinguishable from a person.
4. It aids in providing helpful guidance to its users and may even lead to self-discovery.
5. All personal data is completely secure and never shared outside of the device and its interface.
6. The user is completely confident that point 5 is true and has no worry about data being shared or leaked.

**Questions 2**

1. Given that iSoph does not share data, does it present any privacy harms / risks and if so, what types of harms are they? For example, traditional privacy harms include many classified as psychological, including embarrassment and shame, inhibition, and others. Do these or similar risks arise in this case?
2. Are there potential privacy benefits relevant to the case?
3. How do the concerns about creating systems that imitate conscious humans relate to issues of privacy?
4. What design choices affect the answers to 1-3?

**Exercises 3**

1. Imagine that you are on the design team encharged with specifying the functionality and features of the system. From a design perspective, consider what safeguards could be built into such a conversational agent? For example, would you want to provide options for turning off some features such as realistic voice simulations, or controlling the level of voice realism by degrees or levels?
2. Suppose that you are part of the risk management and governance team performing a privacy impact assessment on the initial design as created by the design team. Would you see the potential benefits sufficient to justify its creation and potential risks of privacy harms? How would you view the safeguards identified by the design team and included in their design plans? Would you consider them sufficient? Are there any features of functions of the system that you would think should simply not be included in the built system?
3. Imagine that you are part of the quality control team and you are asked to include some findings from the risk management team’s analysis in your testing plan. How might you incorporate such findings in your testing plan. For example, if a feature to control how realistic iSoph’s voice were included, how would you test it? Would you simply test its functionality, or would you include user groups in your test plan and observe the effectiveness of the control in terms of its effect on them?
4. Imagine that you are a user of the system and hope to derive benefit from it but are concerned to avoid any negative effects as indicated in user instructions you read. What steps might you take to mitigate or avoid harms. Would you limit your time using it or would you use it for specific purposes only?
5. Suppose you worked for a regulatory agency or legislative body. Are there some potential capabilities of iSoph that you would consider problematic from a privacy perspective? For example, would you want to limit the extent that iSoph could assume roles such as friend or advisor or take steps to form a relationship (e.g., by referring back to past conversations or expressing concern)?
6. As a member of a regulating body, would you want to impose age limits or require caution when used by minors?

**Applying the Principles of AI Ethics 4**

Using the chart provided, identify which principles of AI ethics are at issue in this case and, if principles conflict, which seems to be the weightiest and so the one that should override other principles.

Principle	Application (If Any)
-----------	----------------------

Nonmaleficence	
Beneficence	
Respect for Autonomy	
Justice	
Explicability	
Accountability	

## Normative Theories 5

Apply the different normative theories explained in the primer at the beginning of this volume to bring out issues in the case that might have been overlooked. How would a utilitarian approach this case? A deontologist? A virtue ethicist? Do any of these approaches accord with your own moral judgments or not? Do they arrive at the same verdict as the principles of AI ethics?

## Expert Analysis 6

The question under analysis is whether the personal advisor system, iSoph, raises risks of privacy harms (privacy violations). In addressing the question, the summary points above should be borne in mind. Points 5 and 6 constitute constraints on the case. To rephrase, (P5) iSoph does not share its user's personal data with third parties and it cannot be improperly accessed (hacked) and (P6) the user is certain of P5.

## Description of AI Principles and Privacy Harms 6.1

The analysis will use a principles-based approach to ethics, which is common in applied ethics. The AI ethics principles are an example of this approach and are based in well-established principles in human subject research (Floridi, 2023; Belmont Report<sup>2</sup>). Mooradian (2018) characterizes a principles-based approach as a continuum from high-level principles (e.g., principles of non-maleficence or autonomy) to general moral rules (e.g., non-deception), to domain specific rules, to specific ethical judgments). Domain specific rules are worked out over time as new social phenomena, practices, institutions, and technologies arise. Information privacy and the ethical management of personal information can be thought of as a domain specific ethics with principles<sup>3</sup> codified by international bodies such as the OECD privacy principles.<sup>4</sup> These principles are:

1. Collection Limitation Principle
2. Data Quality Principle
3. Purpose Specification Principle
4. Use Limitation Principle
5. Security Safeguards Principle
6. Openness Principle
7. Individual Participation Principle
8. Accountability Principle

Each of these principles enjoins / prohibits certain behaviors in relation to personal information. For example, the Collection Limitation principle enjoins that the collection of information by organizations should be limited to what is necessary for business purposes. These privacy principles can be viewed as specifications of high-level principles such as the principles of non-maleficence and autonomy. The connection is made in identifying the values and interests that these principles support. We can refer to these as privacy values. Privacy scholars have identified many privacy values and harms. (Solove, 2008) Mooradian (2018) lists a set of common values enabled by privacy:

<ul style="list-style-type: none"> <li>▪ Financial well-being</li> <li>▪ Psychological well-being</li> <li>▪ Freedom of Thought</li> <li>▪ Self-development</li> <li>▪ Individuality</li> <li>▪ Independence</li> <li>▪ Freedom from discrimination</li> </ul>	<ul style="list-style-type: none"> <li>▪ Liberty</li> <li>▪ Autonomy</li> <li>▪ Political participation</li> <li>▪ Dignity</li> <li>▪ Reputation</li> <li>▪ Human relationships</li> <li>▪ Friendship</li> </ul>
--	--

These values are stated in very general terms but are given more specific expression in particular contexts. For example, financial well-being is protected by taking appropriate means to prevent identify theft. Ethical rules for the management of personal information, such as the OECD principles and laws and regulations (e.g., the GDPR), aim to promote these privacy values and avoid privacy harms. They connect directly to the general AI ethical principles. For example, stealing a person's identity and making purchases with her credit cards violates the principle of non-maleficence. It does so for various reasons, including by causing financial harm, but also because it causes mental suffering.

To analyze the case of the iSoph system, we can ask whether any privacy values such as those listed above are potentially infringed upon or undermined. In other terms, we can ask whether iSoph raises the risk of

<sup>2</sup> [The Belmont Report | HHS.gov](https://www.hhs.gov/belmont-report/)

<sup>3</sup> AI ethics can be considered a domain specific ethics that is being developed at a rapid pace as new AI technologies are created and deployed.

<sup>4</sup> [OECD Privacy Principles](https://www.oecd.org/privacy/)

causing any privacy harms. We can also ask whether iSoph may promote any of these values. Based on these answers, we can consider design choices that will minimize privacy harms while maximizing privacy benefits.

Because case summary points 5 and 6 are stipulated by the case, any privacy harms that involve the sharing of personal information (properly or improperly) are ruled out. This covers the vast majority of privacy harms related to the OECD principles and privacy values listed above. For example, identity theft is ruled out as a potential privacy harm because it involves improperly acquiring and using victims' personal information. Identity theft falls within the category of financial harms, as it causes financial damage. It also normally causes mental suffering in the form of worry, frustration, anger, and sadness. As with identity theft, many privacy harms and norms established to prevent them require the (improper) transfer of personal information to another party.

While the majority of privacy harms involve transfer of personal information to a third party or system, there are some significant ones that do not. These harms normally involve direct observation or monitoring of data. Surveillance is an activity that can lead to various privacy harms (Solove, 2008). Knowledge that one is or might be surveilled can cause mental suffering in the form of anxiety. It can also cause one to feel inhibited (mental suffering), which can lead to self-censure and restriction of activities. Such inhibition undermines multiple privacy values such as autonomy and self-development. While concerns about surveillance often include the way the information obtained may be used and/or shared, usage and sharing are not necessary to provoke unease. The mere fact of someone monitoring private thoughts as they are expressed (including via an information technology) or behaviors carried out in private is often sufficient to cause distress and inhibition. For example, a person being viewed in a private space via a one-way mirror would feel his/her privacy is invaded and would feel unease and inhibition. This would be true even if the viewing was not constant, but periodic (the panopticon effect <sup>5</sup>).

Given that the majority of privacy harms are ruled out in relation to the iSoph personal advisor case because data cannot be shared (by stipulation), the question arises as to whether other types of privacy harms such as those caused by direct surveillance are a possibility. Can iSoph, by virtue of its collection of personal information, its direct monitoring via camera/audio, and its ability to learn from data captured, stand in relation to its users as someone surveilling them. In this way, we may have the conditions of a privacy harm. The obvious obstacle for iSoph to stand in the role of one surveilling / monitoring another is that iSoph is an AI system, not a person. As an AI system, it exceeds persons in certain surveillance capabilities. However, as an AI system, iSoph is not a conscious human socialized within the community of the user. iSoph is not a person with a personality. So, it would seem that iSoph would not be capable of causing privacy harms based in surveillance.

## Ethics of Simulating Sentience 6.2

iSoph's lack of personhood would seem to disqualify it from causing privacy harms through direct observation of its users. However, a significant area of artificial intelligence research is dedicated to creating systems that simulate sentient or even conscious beings including socialized human persons (Donath, 2020). These include social robots used to provide companionship to residents in assisted living and conversational agents used in customer service. Recent advancements in large language models (e.g., ChatGPT) have resulted in systems that produce fluent conversation that, in various contexts, is indistinguishable from human-created text. As a result, persons interacting with such systems often behave as if they were interacting with another person, even when they know they are conversing with an AI system. According to numerous researchers, our tendency to act as if an artificial system is a human interlocutor is based in the linguistic competencies required for us to carry out conversations with human speakers and to read text produced by persons. As Bender et al. state, ". . . our perception of natural language text, regardless of how it was generated, is mediated by our own linguistic competence and our predisposition to interpret communicative acts as conveying coherent meaning and intent, whether or not they do" (2021, p. 616). In the case of AI systems, they do not. They are non-conscious conversational agents using probabilistic algorithms to generate realistic speech without having and sharing a mental model of the world with us. Our imputation of a shared understanding of our context is an illusion the

---

<sup>5</sup> The panopticon concept was proposed by the philosopher Jeremy Bentham. It is a prison architecture that allows guards to view any prison cells at any time, but not all at once. Prisoners do not know when they are being watched.

creation of which is an ethical concern in itself (Ibid.) As Donath argues, a system designed to encourage such illusions is inherently deceptive.

Identity deception of some kind is inherent to all artificial, seemingly sentient entities: they are made to look, act, and/or speak as if a thinking, feeling, sensing mind was motivating them. Even for one to declare "I am a program" is, arguably, deceptive, for the use of the word "I" implies a thinking self-aware existence, the being whose thought process formed those words (2020, p. 69)

In online contexts, users are often unaware that they are conversing with an artificial agent. Here the deception is complete. To accomplish such indistinguishability from human interlocutors, it helps greatly that contexts of interaction are constrained by subject matter and duration. Longer, open ended conversations pose a greater challenge to passing as human. Even so, systems that simulate human-like conversation achieve a partial illusion because the linguistic and social capabilities they engage operate (in large part) on a subconscious level. Just as we do not consciously attend to questions of grammar when speaking (unless we run into a problem), so we do not consciously monitor our attributions of shared mental models (unless, again, we run into a problem, for example, a failure to make sense of the other person's speech). As a result, AI systems designed to engage these capabilities can do so successfully, even when the person interacting with the system knows that it is a computer system. Further, when linguistic and social dispositions are engaged, social behaviors and emotions may also be engaged. Like speech, social behavior and emotional response is learned via membership in a community and it operates in large part at a subconscious level. For this reason, it is possible to be disposed to be polite to an AI conversational agent when engaging in a transaction, and it is also possible to feel annoyance or disturbance if the AI agent is impolite to you.

### Interpersonal Privacy Harms and Conversational Agents 6.3

If the above is correct, it is possible to see how iSoph could give rise to privacy harms, in particular, harms associated with surveillance. iSoph's ability to function as a surveillance system was described in the above section on AI Principles and Privacy Harms. What was lacking in the description was the presence of a conscious entity with a similar socialization to ours, to wit, another person or persons. The previous section on the Ethics of Simulating Sentience described how non-conscious entities can stand in for human interlocutors and can evoke emotional responses from persons interacting with them. Advanced systems designed to produce coherent speech in a way that simulates human speech can trigger linguistic and emotional responses from persons interacting with the system, even when they "know" they are dealing with a system. So designed, an AI system presents a kind of "reverse Turing test" in which the goal is to have its users respond to it (to some extent) as another person, even when it is known in advance to be a computer system.<sup>6</sup> Practical reasons for engaging emotional responses include providing customer service interactions via conversational agents that feel satisfactory to customers who would presumably prefer to deal with an actual person. For the purposes of this discussion, the important points are (a) human speech and conscious behavior can be simulated by AI conversational agents, and (b) persons interacting with such systems can half-believe that they are interacting with persons to the extent that their linguistic and emotional dispositions are triggered.

Putting these points together, we can see that the iSoph system, as described, could generate some of the privacy harms that arise from surveillance (that is, constant observation). First, it has extensive knowledge gained from observation, data collection, and inferential ability. It therefore holds and continues to generate a vast amount of personal information. Some of this information is sensitive and some might be described as secret. Some of this information is not previously known by the user. Second, iSoph is designed to interact verbally with its user in a way that is practically indistinguishable from normal human verbal behavior. It is therefore capable of interacting with its user in a way that evokes linguistic and emotional responses of the sort that human speech and interaction evoke.

To see how iSoph's capabilities could give rise to privacy harms, consider the following interaction. One afternoon, while our user is browsing social media sites, iSoph comments in a human-simulated voice that s/he

---

<sup>6</sup> This is the plot line of the science fiction movie *Ex Libris*.

has been spending a great deal of time engaged in such activities (a precise number is given in hours and minutes), that s/he should spend more time reading articles or books, and that s/he should read a set of articles on politics that would serve to balance his/her perspective (a list of recommendations is produced). We can then imagine that iSoph begins at some point making such comments frequently on a wide range of observed behaviors and topics. Some of these comments, like the ones just cited, might be useful and even welcomed. Others, however, may not. Further, the cumulative effect of such comments may lead our user to feel the s/he is constantly being watched, and even judged. The helpful comments may “feel” like criticism, and the feeling of being criticized may lead to the feeling of being constantly observed and even judged. As a result, our user may feel discomfort or anxiety, both of which are forms of mental suffering, and inhibition, a form of behavior that negatively impacts personal autonomy. If this is the case, a number of privacy harms as listed above will have been caused.

The risk of these privacy harms can be understood within the AI ethical framework. iSoph could potentially cause emotional distress in the form of anxiety, embarrassment, insecurity, and other similar emotions. These may all rise to the level of mental suffering, especially when they occur on a continuing basis. Mental suffering is a form of harm that falls under the principle of non-maleficence. Therefore, an explanation of its wrongfulness can be grounded in this principle. Similarly, the principle of autonomy can be applied to the case, as the feelings of inhibition that arise from being watched in a private space can lead to self-censoring and curtailment of certain behaviors. This principle can therefore be used to ground an explanation of the risks presented by iSoph. Other AI principles may also be explored in relation to the case. (This is left to the reader as an exercise.) It is, in fact, common that multiple principles apply to a given privacy harm given the role that privacy plays in supporting a multiplicity of values.

#### Privacy Benefits 6.4

The above section provides a description of how iSoph may raise privacy concerns, despite the fact that it does not share data with third parties. It situates the privacy risks within a familiar set of values enabled by privacy. Since privacy harms are understood in terms of their negative relation to privacy values, we can also ask whether and how iSoph might also promote privacy values. It is easy to imagine that, as described in our case, iSoph could bring about enjoyment by providing helpful and interesting information, recommendations, and advice. These interactions could be enjoyable in themselves in many different ways but, more importantly, the information and guidance could lead to choices that lead to long term mental well-being and flourishing. For example, insights provided into sources of suffering could lead to better decisions, changes in behavior, and fruitful information seeking.<sup>7</sup> Further, the provision of personally relevant information would likely strengthen a user’s general knowledge and self-knowledge. This, in turn, would promote his or her personal autonomy. It would help with making specific decisions as well as exploring different conceptions of how to live her life. So, just as iSoph may pose risks in the areas of mental suffering and autonomy, it could also promote psychological well-being and autonomy. It would likely promote other privacy enabling values as well, e.g., financial well-being.

#### Ethical Design 6.5

Given that iSoph has the potential to cause privacy harms, but also the potential to create privacy benefits, any ethical analysis of the system as built and deployed would need to take both harms and benefits into account. But in the context of AI systems (and software generally), ethical analysis should not wait until a system is built. Rather, it should be a factor in the design of the system. This is the concept of ethical design, which is based in or inspired by privacy by design.<sup>8</sup> From an ethics-by-design perspective, ethical considerations should be taken as requirements for a to-be-designed system, just like other system requirements. That is, they should be included in the goals of the system and be implemented in its functionality and operational principles.

---

<sup>7</sup> Not coincidentally a long-standing use case for conversational AI going back to Weizenbaum’s ELIZA system has been psychological therapy.

<sup>8</sup> Privacy by Design is the concept that privacy features should be included in the design of a system.




If we take an ethical design perspective, we will examine which features in iSoph are ethically salient, which is to say, which features pose ethical risks and which carry benefits. We will aim to minimize risks and maximize benefits. The benefits of the iSoph system, as described, consist in its ability to provide valuable information in the form of personal advice. The specific privacy risks we have identified are attributable to its ability to simulate human speech and trigger our social-linguistic capabilities in ways we do not fully control. An ethical design will therefore retain the beneficial informational capabilities, while eliminating or making optional the human-imitative capabilities. Interestingly, it is the human-imitative capabilities that are the center of concern for many authors writing on the topic of AI ethics. As quoted above, Bender et al. and Donath raised concerns about the inherently deceptive nature of simulated human speech. So, our analysis serves to add another risk area to ones already identified by these authors and others.

To mitigate the risks created by human-imitative speech capabilities, developers can use other interface/output types. An interface based in a search paradigm with more structured information, including linked data, would serve to provide a rich informational interface that does not trigger interpersonal emotional responses. It could even have a voice option as do text-to-speech readers in document applications such as Adobe Acrobat or MS Word. The form of such speech, however, would be deliberately non-conversational. Also, the interface, being based in a search paradigm, would wait for questions or commands as input as opposed to being allowed to freely comment or initiate conversation. Other design ideas such as these could be incorporated to avoid simulative effects. In this way, iSoph would function as a powerful, beneficial information system that maximizes a number of values, including privacy values, while minimizing harms. An ethical design perspective, in this case, will serve to make the system more beneficial and effective in meeting its objectives. Additionally, it will illustrate the value of ethical analysis as a component of system design.

## Student Reflection 7

Did you touch on everything this expert analysis identifies in your own analysis of the case? What did you miss? Did you think of anything that could be added to the analysis or were there any points of disagreement? Be sure to justify your response with reasons.

## References

- Bender, E., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ACM, 2021.
- Donath, J. "Ethical Issues in Our Relationship with Artificial Entities." *The Oxford Handbook of Ethics of AI*, by Judith Donath, edited by Markus D. Dubber et al., Oxford University Press, 2020, pp. 51–73.
- Floridi, Luciano. *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford University Press, 2023.
- Gabriel, I. "The Ethics of Advanced AI Assistants." Google DeepMind. 2024.
- Mooradian, N. *Ethics for Records and Information Management. First Edition*, ALA Neal-Schuman, an imprint of the American Library Association, 2018.
- Solove, D. *Understanding Privacy*. Harvard University Press, 2008.